



SAPIENZA
UNIVERSITÀ DI ROMA

Multimodal Knowledge-Intensive NLP: From Representation to Augmentation and Evaluation

Dipartimento di Ingegneria Informatica, Automatica e Gestionale "Antonio Ruberti" (DIAG)

Dottorato Nazionale in Intelligenza Artificiale (XXXVIII cycle)

Karim Kamal Ghonim

ID number 1774086

Advisor

Prof. Roberto Navigli

Academic Year 2025/2026

Thesis defended on – May 2026
in front of a Board of Examiners composed by:
Prof. Emanuele Frontoni (chairman)
Prof. Danilo Comminiello
Prof. Cristian Randieri

This thesis was reviewed by two external reviewers:
Prof. Raffaella Bernardi
Prof. Sandro Pezzelle

Multimodal Knowledge-Intensive NLP: From Representation to Augmentation and Evaluation
PhD thesis. Sapienza University of Rome

© 2026 Karim Kamal Ghonim. All rights reserved

This thesis has been typeset by $\text{\LaTeX}$ and the Sapthesis class.

Author's email: ghonim@diag.uniroma1.it

Dedicated to my friends and family ...

Abstract

Modern Artificial Intelligence (AI) and Natural Language Processing (NLP) systems are increasingly deployed in knowledge-intensive scenarios that require broad and up-to-date world knowledge, ranging from question answering and information retrieval to digital assistants and automated fact-checking. This progress is largely driven by pretrained Large Language Models (LLMs) and multimodal models, which acquire substantial information through training on large-scale corpora. Yet, despite these advances, current models still face three recurring limitations that hinder reliable deployment.

First, their knowledge of concepts and entities is uneven and often skewed toward the domains and distributions represented in widely used training resources and benchmarks. Second, they are typically deployed under a fixed-snapshot view of world knowledge, mirroring the static nature of the inventories that support many downstream pipelines, such as knowledge bases, which makes it difficult to handle newly emerging entities and rapidly evolving facts. Third, when models generate content in settings where new entities and information emerge continuously, current evaluation practices provide limited assurance of factual accuracy and little transparency into which statements are unsupported and why. These challenges become even more pronounced in multimodal settings, where systems must jointly leverage visual and textual evidence to access and validate knowledge.

Motivated by these shortcomings, this thesis advances knowledge-intensive language and multimodal AI along three complementary directions that jointly support coverage, dynamic augmentation, and verification. We first develop large-scale resources and evaluation protocols that broaden semantic coverage beyond the narrow scope of widely used datasets, enabling systematic analysis of representation gaps and supporting the development and assessment of models that generalize across diverse concepts and entities. However, even broad-coverage static resources cannot keep pace with knowledge that emerges after their

construction. To address this, we introduce retrieval-augmented generation methods that leverage externally retrieved evidence to produce factually grounded information for newly emerging entities, enabling knowledge inventories to be updated over time without retraining the downstream models that rely on them. As automatically generated content becomes a pathway for continual knowledge integration, ensuring its reliability becomes essential. We therefore propose evidence-grounded evaluation frameworks that decompose model outputs into verifiable units and validate each against source documents, providing fine-grained, interpretable factuality assessment suitable for reference-scarce settings.

Overall, this thesis advances AI systems that (i) cover a broader portion of real-world semantics, (ii) remain effective as knowledge evolves and new entities appear, and (iii) can be assessed with transparent, evidence-based factuality signals, enabling more reliable deployment in knowledge-intensive applications.

~ This page was intentionally left blank ~

Contents

1	Introduction	1
1.1	Overview	1
1.2	Knowledge in a Knowledge-Intensive World	2
1.3	Challenges in Knowledge-Intensive AI	4
1.4	Thesis Statement and Research Objectives	6
1.5	Thesis Outline	7
1.6	Publications and Contributions	9
2	Background and Related Work	11
2.1	Knowledge Resources	11
2.1.1	Wikipedia: Encyclopedic Knowledge for Concepts and Entities . .	12
2.1.2	WordNet: Lexical Taxonomies for Concepts	14
2.1.3	ImageNet: Visual Taxonomies Grounded in WordNet	16
2.1.4	BabelNet: Unifying Lexicographic and Encyclopedic Knowledge .	20
2.1.5	CNER: Coarse-Grained Semantic Categories for Concepts and Entities	22
2.2	Knowledge-Intensive Multimodal Tasks	23
2.2.1	Visual Concept Recognition	23
2.2.2	Multimodal Entity Linking	29
2.2.3	Entity Linking with Emerging Entities	31
2.3	Dynamic Knowledge Integration	35
2.3.1	Knowledge Editing and Model Updating	35
2.3.2	Retrieval-Augmented Modeling	36
2.3.3	Knowledge Integration under Dynamic Knowledge	38
2.4	Knowledge-Intensive Evaluation	39

2.4.1	Limitations of Reference-Based Similarity Metrics	39
2.4.2	Factual Consistency and Evidence-Grounded Evaluation	39
2.4.3	Evaluation Under Evolving Knowledge	41
3	Addressing the Gap in Multimodal Knowledge Coverage	45
3.1	Overview	45
3.2	Vision-Language Concept Recognition	46
3.3	Concept-pedia	48
3.3.1	Link propagation	49
3.3.2	Taxonomical Concept Propagation	51
3.3.3	Dataset Statistics	53
3.4	Concept-10k	53
3.4.1	Benchmark Statistics and Analysis	54
3.5	Dataset Evaluation	56
3.5.1	Annotation Quality Assessment	57
3.5.2	Comparison with ImageNet-1k	58
3.5.3	Benchmark Coverage Comparison	59
3.5.4	Annotation Guidelines	60
3.6	Experiments and Results	65
3.6.1	Experimental setup	65
3.6.2	Zero-shot Visual Concept Recognition	67
3.6.3	Concept-pedia Fine-tuning	68
3.7	Summary	70
4	Large-Scale Multimodal Knowledge - Beyond Concepts	73
4.1	Overview	73
4.1.1	From Concepts to Entities	74
4.2	Multimodal Entity Linking	75
4.2.1	Challenges and Limitations	75
4.3	Enti-pedia: A Large-Scale Resource for Multimodal Entity Linking	76
4.3.1	Enti-pedia Scale and Coverage	78
4.4	Experimental Setup	79

4.4.1	Evaluation Benchmarks	79
4.4.2	Models and Baselines	79
4.4.3	Enti-pedia pre-training	80
4.4.4	Generalization Evaluation in EL	81
4.5	Results and Discussion	82
4.5.1	In-domain versus Cross-dataset Evaluation	82
4.5.2	Results	84
4.5.3	Discussion	85
4.6	Summary	86
5	Emerging Knowledge Augmentation	89
5.1	Overview	89
5.2	Static Knowledge and Emerging Entities	91
5.2.1	Challenges and Limitations	92
5.3	RAED	94
5.3.1	Passage Index	94
5.3.2	Retriever	95
5.3.3	Entity Description Generation	96
5.3.4	RAED for Entity Disambiguation	96
5.4	Experimental Setup	97
5.4.1	Retriever	97
5.4.2	Entity Description Generation	97
5.4.3	Datasets	98
5.4.4	Evaluation Metrics	99
5.5	Results	100
5.5.1	Entity Description Generation	101
5.5.2	Entity Disambiguation	102
5.5.3	Emerging Entity Linking	103
5.5.4	Qualitative Analysis	104
5.6	Summary	106

6	Toward Interpretable Knowledge Evaluation	109
6.1	Overview	109
6.2	Factuality Evaluation in Knowledge-Intensive Generation	111
6.3	FENICE	114
6.3.1	Claim extraction	114
6.3.2	NLI-based claim scoring	114
6.3.3	Coreference Resolution	116
6.3.4	Aligning claims across multiple input text granularities	117
6.4	Experiments and Results	118
6.4.1	Claim extraction	118
6.4.2	Factuality Evaluation	120
6.4.3	Ablation study	122
6.4.4	Long-form summarization evaluation	124
6.5	FENICE for Interpretable Entity Description Evaluation	127
6.6	Summary	129
7	Conclusions and Future Work	133
7.1	Overview	133
7.2	Summary of Contributions	134
7.3	Future Work	136
	Bibliography	139

Chapter 1

Introduction

1.1 Overview

In recent years, Artificial Intelligence (AI) has undergone rapid progress, driven largely by the emergence of large-scale pretrained language models and foundation models capable of performing a wide range of tasks with minimal task-specific supervision (Brown et al., 2020; Bommasani, 2021; Touvron et al., 2023). These models exhibit strong capabilities in language understanding, reasoning, and generation, and increasingly form the backbone of real-world systems such as conversational agents, search engines, programming assistants, and decision-support tools.

Although early breakthroughs in modern AI focused predominantly on text, recent advances have expanded these systems to jointly process multiple modalities, including images, audio, and video. Vision-language foundation models now integrate visual signals with linguistic information, allowing systems to reason over heterogeneous inputs in ways that more closely resemble how humans perceive and interpret the world (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023a; Team et al., 2025). This shift toward multimodal processing has accelerated the deployment of AI in everyday environments, where users increasingly rely on such systems to interpret complex inputs and support real-world decision-making.

As these systems become more deeply integrated into daily life, the role of the knowledge they retain becomes central. Users increasingly expect AI systems to provide factually consistent information, resolve ambiguities, and make coherent, knowledge-grounded judg-

ments about real-world phenomena. Meeting these expectations fundamentally depends on access to accurate world knowledge, as reliable behavior requires models to access and apply knowledge about entities, concepts, and events. However, despite recent advances, knowledge remains a major bottleneck for current models. Much of a model’s knowledge is stored implicitly in its parameters, rendering it difficult to inspect and largely static once training is complete. Consequently, it is often unclear whether a model’s outputs reflect grounding in reliable real-world information or spurious correlations acquired during pre-training (Petroni et al., 2019; Roberts et al., 2020; Kandpal et al., 2023). Furthermore, parametric knowledge does not readily adapt to newly emerging information, such as newly introduced entities or rapidly evolving facts, without retraining.

To better understand and quantify how models access and use knowledge, the research community has increasingly turned to *knowledge-intensive* tasks. These tasks are explicitly designed to require access to a large body of information in order to perform correctly, analogous to the way humans consult documents, encyclopedias, and databases when answering factual questions or resolving ambiguities. Performance on such tasks therefore provides a principled basis for evaluating both the knowledge encoded within models and their ability to retrieve, integrate, and reason over external evidence when needed (Petroni et al., 2021; Guu et al., 2020; Lewis et al., 2020b).

However, real-world knowledge is neither homogeneous nor static. It spans multiple forms and levels of abstraction, from visual concepts and semantic categories to uniquely identifiable entities and detailed factual assertions. Moreover, the world is continually evolving: new entities emerge, existing facts change, and previously accepted information may be updated or revised. Capturing this diversity and temporal dynamics remains challenging for current modeling and evaluation paradigms, which often focus on narrow domains and fixed benchmarks. Consequently, understanding how models represent, extend, and validate knowledge across these dimensions is essential for building reliable AI systems.

1.2 Knowledge in a Knowledge-Intensive World

Knowledge refers to the information about the world that enables an intelligent system to interpret observations, resolve ambiguity, and make informed decisions. For humans, this

knowledge is acquired through experience, education, and access to external resources such as books, media, and collective memory. When encountering an unfamiliar term, a historical event, or a visual object, people naturally consult references, recall prior knowledge, and integrate multiple sources of evidence before forming a conclusion.

In modern AI systems, the acquisition and use of world knowledge are increasingly mediated by large pretrained models and access to external information sources. A language model may implicitly encode facts about public figures, locations, or scientific concepts within its parameters, or draw upon external documents and structured resources available at inference time. However, these systems lack an explicit representation of the information they retain and clear mechanisms for tracing why a particular output is produced.

A task is typically considered *knowledge-intensive* when correct behavior cannot be achieved through linguistic fluency alone, but instead requires access to factual information beyond the input itself. For example, answering the question “Who directed *Inception*?” requires knowledge about a specific movie and its director; resolving which president a mention of “The President” refers to in a given text depends on contextual world knowledge; and identifying a landmark in an image may require recognizing a particular building and linking it to a named entity. In all these cases, success depends on the availability and correct use of knowledge that is either internalized within the model or retrieved from external sources.

More broadly, real-world knowledge is not uniform in structure or granularity. It spans multiple levels of abstraction and appears in different representational forms. At a coarse level, which we refer to as *conceptual knowledge*, systems must recognize general concepts and categories, such as distinguishing animals from vehicles or identifying visual properties like color and texture. At a more fine-grained level, referred to as *entity-level knowledge*, systems must reason about concrete instances: individual people, organizations, locations, products, or events that are uniquely identifiable and grounded in knowledge bases. Beyond discrete labels, much knowledge is expressed in unstructured form through natural language descriptions, narratives, and documents that encode relations, properties, and temporal evolution. These forms of knowledge play complementary roles. Conceptual knowledge supports abstraction and generalization, enabling transfer across instances and domains, while entity-level knowledge anchors reasoning to specific real-world referents,

supporting precise identification and disambiguation.

Furthermore, real-world knowledge is inherently dynamic. New entities emerge continuously, existing entities evolve, and factual information changes as the world itself changes. Static training corpora and fixed benchmarks therefore capture only a snapshot of reality in a specific moment in time. As models are increasingly deployed in open environments, the ability to integrate new evidence, update internal representations, and validate generated information becomes as important as the breadth of knowledge acquired during training.

The properties of real-world knowledge discussed above highlight three recurring challenges for knowledge-intensive AI systems: achieving broad and reliable knowledge coverage across multiple levels of abstraction, enabling dynamic integration of emerging information, and validating generated content in a grounded and interpretable manner.

1.3 Challenges in Knowledge-Intensive AI

Knowledge Coverage. Many knowledge-intensive tasks, and particularly multimodal ones, implicitly assume that models have access to broad, reliable world knowledge. However, existing resources and benchmarks, while often large in size, exhibit narrow semantic and topical coverage, focusing on a limited set of domains and commonly used distributions. For example, standard visual datasets and multimodal benchmarks are often dominated by concrete object categories or coarse-grained concepts (Deng et al., 2009; Ridnik et al., 2021; Horn et al., 2018). This narrow coverage constrains both model development and evaluation, leading to systems that perform well in familiar domains but exhibit performance degradation when confronted with less frequent, fine-grained, or out-of-distribution knowledge, particularly when multiple modalities must be integrated (Wang et al., 2022c; Yang et al., 2023).

Emerging Knowledge. Several knowledge-intensive tasks are grounded in knowledge bases that represent a snapshot of the world at a particular point in time and are treated as sources of ground truth. A prominent example is Entity Linking (EL), which aims to map entity mentions in text to corresponding entries in a knowledge base. Many EL models operate under a closed-world assumption in which the underlying knowledge base and

candidate entity set are treated as static and complete. Traditional EL frameworks condition their predictions on a fixed inventory of entities and learn to resolve mentions only to those present in a predefined set; this design fundamentally limits their applicability in real-world settings, where new entities continuously emerge and existing knowledge evolves over time. In dynamic information environments, such as news streams and scientific publications, novel concepts, organizations, and individuals appear on a daily basis, rendering static knowledge inventories insufficient for practical deployment without frequent retraining. Because most EL systems are not explicitly designed to incorporate emerging information dynamically, their performance can degrade substantially when evaluated on temporally evolving or open-world settings that contain previously unseen entities (Zaporojets et al., 2022).

Interpretable Knowledge Evaluation. As AI systems increasingly handle incomplete or evolving information, the ability to evaluate model outputs in an automated, reliable, and explainable manner becomes essential for practical deployment. Because real-world knowledge is inherently dynamic, gold references may be unavailable, incomplete, or quickly outdated for validating model outputs. Evaluation therefore can no longer rely solely on direct comparison against predefined targets, but must instead determine whether an output is supported by available evidence. In such settings, traditional evaluation approaches based on surface overlap, semantic similarity, or reference matching provide limited insight into factual correctness and offer little transparency into why an output should be considered reliable (Kryscinski et al., 2020; Durmus et al., 2020; Honovich et al., 2021; Laban et al., 2022). As a result, factual inconsistencies may remain undetected and propagate to downstream systems, undermining trust in models designed to operate over diverse and evolving information. Addressing this challenge requires evaluation frameworks that explicitly validate model outputs against supporting evidence and provide interpretable signals about their reliability. Such mechanisms are essential for deploying systems that must operate over broad knowledge under continual change.

1.4 Thesis Statement and Research Objectives

This thesis aims to improve the reliability of knowledge-intensive multimodal systems by addressing three core requirements: broad and robust knowledge coverage, dynamic integration of emerging information, and interpretable validation of generated knowledge. More specifically, the goal is to move beyond narrow-coverage benchmarks, dataset-specific evaluation settings, and static knowledge assumptions. This shift is necessary to enable systems that can operate robustly in real-world environments, where knowledge is diverse, continuously evolving, and must be grounded in verifiable evidence. The main objectives that this thesis aims to accomplish are outlined as follows:

1. **Expanding Multimodal Knowledge Coverage.** Current multimodal resources and evaluation benchmarks capture only a limited subset of real-world semantic diversity, often concentrating on a narrow concept and entity distributions. A primary objective of this thesis is to construct large-scale multimodal resources and benchmarks that substantially broaden semantic coverage at both the concept and entity levels. This enables the systematic advancement and analysis of model generalization across diverse semantic categories and reduces reliance on benchmark-specific conclusions.
2. **Establishing Scalable Multimodal Entity Understanding.** Beyond broad conceptual coverage, Knowledge-intensive applications require the identification and disambiguation of concrete real-world entities rather than merely recognizing abstract categories. To this end, we establish large-scale multimodal resources and training paradigms that enable robust Multimodal Entity Linking (MEL), improve cross-dataset generalization, and mitigate overfitting to narrow domain distributions, thus improving scalability to open-domain and real-world settings.
3. **Enabling Dynamic Integration of Emerging Knowledge.** Even when supported by large-scale resources, current systems remain constrained by the assumption that knowledge inventories are fixed at inference time. This constraint prevents models from handling newly emerging entities without retraining or rebuilding the underlying inventories they rely on. A central objective of this thesis is to address this limitation by automatically generating factual, evidence-grounded knowledge for emerging

entities using externally retrieved sources. This generated knowledge can be incorporated into existing inventories to keep them up to date without human intervention, thereby enabling downstream tasks such as Entity Linking in continuously evolving environments.

4. **Interpretable and Evidence-Grounded Knowledge Validation.** Emerging knowledge is inherently difficult to evaluate because, in realistic settings, gold reference descriptions may be unavailable, incomplete, or quickly outdated. To enable reliable use of automatically generated updates, we develop an interpretable factuality evaluation framework that verifies model outputs directly against supporting evidence. Through qualitative and quantitative analysis, we empirically show that this framework provides fine-grained validation signals, supporting the filtering and auditing of automatically generated knowledge before its integration into knowledge inventories.

The overarching goal of this thesis is to enable knowledge-intensive AI systems that achieve broad and robust knowledge coverage, dynamically integrate emerging and continuously evolving knowledge, and produce factual and verifiable outputs suitable for reliable deployment in real-world environments.

1.5 Thesis Outline

This thesis examines how modern NLP and multimodal systems can represent, extend, and validate world knowledge in realistic settings, where knowledge is diverse in form, continuously evolving, and only partially observable through static training data. As discussed in the introduction, addressing these challenges requires advances across multiple components, including the construction of wide-coverage resources, the development of effective modeling paradigms, and the design of reliable and interpretable evaluation methodologies.

The dissertation is organized to reflect this progression. Rather than following the chronological order of publication, the chapters are structured to guide the reader through the central challenges one at a time, moving from knowledge representation, to dynamic knowledge integration, and finally to knowledge validation.

Chapter 2 introduces the background and related work that ground the thesis. We provide a formal discussion of knowledge-intensive tasks in NLP and multimodal settings,

and review prior work on large-scale knowledge resources, Visual Concept Recognition, Multimodal Entity Linking, retrieval-augmented generation, and factuality evaluation. This chapter establishes the conceptual and methodological context for the contributions that follow.

Chapter 3 addresses the problem of knowledge coverage at the conceptual level. We introduce *Concept-pedia*, a large-scale semantically annotated multimodal resource grounded in Wikipedia, together with *Concept-10k*, a benchmark designed to evaluate Vision-language Models across a substantially broader range of semantic categories than existing datasets. Through extensive analysis, we show that current models exhibit limitations in semantic coverage and generalization beyond narrow benchmark distributions.

Chapter 4 extends the investigation from concepts to entities, focusing on large-scale entity-centric multimodal knowledge. We introduce *Enti-pedia*, a resource for pre-training Multimodal Entity Linking models with broad entity coverage and rich visual and textual supervision. Experimental results show that models trained on Enti-pedia generalize more robustly across datasets and domains than models trained on dataset-specific resources, highlighting the importance of scalable entity-level knowledge representations.

While improved coverage strengthens static representations, real-world knowledge remains inherently dynamic. Chapter 5 addresses the challenge of emerging entities and evolving information. We introduce *RAED*, a model capable of generating entity descriptions by grounding generation in external evidence rather than relying on fixed knowledge base inventories. This approach enables systems to handle previously unseen entities and to incorporate newly available information without requiring costly retraining.

Chapter 6 focuses on the validation of automatically generated knowledge. We introduce *FENICE*, an interpretable factuality evaluation framework that decomposes generated text into atomic claims and verifies them against source documents using natural language inference. We demonstrate how FENICE enables fine-grained, evidence-based analysis of factual consistency and apply it to the evaluation of retrieval-augmented entity description generation, illustrating how document-grounded evaluation supports reliable knowledge integration.

Finally, Chapter 7 summarizes the main findings of the dissertation and discusses broader implications and future research directions, including multilingual knowledge cov-

erage, multimodal emerging entities, learning from factuality signals, and robust evaluation under temporal knowledge drift.

1.6 Publications and Contributions

The research presented in this thesis was conducted at Sapienza NLP, the research group directed by Professor Roberto Navigli, during the three years of the National Ph.D. program in Artificial Intelligence. Most of the work included in this dissertation has been published in top-tier venues in Natural Language Processing and Artificial Intelligence, including ACL, EMNLP, and the Computational Linguistics (CL) Journal.

In this thesis, we discuss in depth only those publications that are directly aligned with the research objectives presented in Chapter 1. In what follows, we provide a list of these publications in reverse chronological order:

1. Ghonim et al. (2025a): **Karim Ghonim**, Andrei Stefan Bejgu, Alberte Fernández-Castro, and Roberto Navigli. 2025. Concept-pedia: a Wide-coverage Semantically-annotated Multimodal Dataset. *In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34417–34438, Suzhou, China. Association for Computational Linguistics.*

Contribution: I proposed the main research idea, designed the dataset construction methodology, supervised the manual annotation process, created the benchmark, implemented the data processing pipeline, carried out all experiments, and led the writing of the paper.

2. Ghonim et al. (2025b): **Karim Ghonim**, Pere-Lluís Huguet Cabot, Riccardo Orlando, and Roberto Navigli. 2025. RAED: Retrieval-Augmented Entity Description Generation for Emerging Entity Linking and Disambiguation. *In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 34439–34452, Suzhou, China. Association for Computational Linguistics.*

Contribution: I proposed the core idea and task formulation, designed the dataset and evaluation protocol, supervised annotation, implemented the retrieval-augmented framework, conducted all experiments, and led the writing of the paper.

3. Scirè et al. (2024b): Alessandro Scirè, **Karim Ghonim**, and Roberto Navigli. 2024. FENICE: Factuality Evaluation of summarization based on Natural language Inference and Claim Extraction. *In Findings of the Association for Computational Linguistics: ACL 2024, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.*

Contribution: I designed and implemented the experimental framework, developed the claim extraction pipeline, conducted the evaluation studies, and contributed to the writing and analysis of the paper.

In addition to the core publications presented in this dissertation, other research efforts carried out during the Ph.D. program have informed the direction of this work. We briefly reference relevant insights and methodologies from the following publications:

4. Scirè et al. (2024a): Alessandro Scirè, Andrei Stefan Bejgu, Simone Tedeschi, **Karim Ghonim**, Federico Martelli, and Roberto Navigli. 2024. Truth or Mirage? Towards End-to-End Factuality Evaluation with LLM-OASIS. *To appear in Computational Linguistics.*

Contribution: I was responsible for the claim extraction component, contributed to the dataset construction and evaluation protocol, implemented parts of the experimental pipeline, and co-authored the paper. I also contributed to the analysis of end-to-end factuality evaluation experiments.

5. Martinez Lorenzo et al. (2024): Abelardo Carlos Martinez Lorenzo, Pere-Lluís Huguet Cabot, **Karim Ghonim**, Lu Xu, Hee-Soo Choi, Alberte Fernández-Castro, and Roberto Navigli. 2024. Mitigating Data Scarcity in Semantic Parsing across Languages with the Multilingual Semantic Layer and its Dataset. *In Findings of the Association for Computational Linguistics: ACL 2024, pages 14056–14080, Bangkok, Thailand. Association for Computational Linguistics.*

Contribution: I was responsible for the Arabic experiments, contributed to the dataset construction and evaluation protocol, implemented parts of the experimental pipeline, and co-authored the paper.

Chapter 2

Background and Related Work

This chapter provides the background needed to situate the thesis contributions and to position our methods relative to prior work. We focus on the resources, tasks, and evaluation protocols that recur in knowledge-intensive multimodal systems, and we highlight how temporal change complicates both modeling and evaluation.

In summary, this chapter covers:

- Knowledge resources for grounding, including encyclopedic and lexical substrates such as Wikipedia and WordNet (Section 2.1);
- Knowledge-intensive multimodal tasks and benchmarks that require grounding across text and images (Section 2.2);
- Knowledge integration under dynamic knowledge, including strategies that reduce reliance on static snapshots (Section 2.3.3);
- Knowledge-intensive evaluation, including factual consistency, evidence attribution, and evaluation under evolving knowledge (Section 2.4).

2.1 Knowledge Resources

A recurring theme of this thesis is that knowledge-intensive settings require models to associate observed inputs, including text, images, or both, with canonical identifiers to denote concepts, entities, and their relations. Achieving this association at scale presupposes

the availability of large-scale knowledge resources that are simultaneously broad in coverage, semantically structured, and continuously maintained.

Historically, knowledge resources in AI have occupied two ends of a spectrum: fully structured repositories, such as manually curated ontologies and knowledge bases, offer high precision but limited coverage, whereas unstructured corpora provide broad coverage at the cost of noisy, implicit semantics. As argued by Hovy et al. (2013), collaboratively built semi-structured resources, most prominently Wikipedia, occupy a productive middle ground: their combination of free-text prose, hyperlink structure, infoboxes, category systems, and inter-language links yields content that is at once semantified, wide-coverage, up-to-date, multilingual, and freely accessible. These properties have made such resources a cornerstone for acquiring computationally exploitable knowledge and deploying it across a wide spectrum of AI and NLP.

In contemporary NLP and multimodal pipelines, the association between observed inputs and world knowledge is typically mediated by a small set of widely adopted resources that provide (i) encyclopedic coverage of real-world concepts and entities (Section 2.1.1), (ii) lexical taxonomies that support conceptual abstraction (Section 2.1.2), and (iii) mappings across resources that enable consistent normalization across languages and modalities (Section 2.1.4). Accordingly, the remainder of this section introduces the core resources that underpin the datasets, benchmarks, and tasks discussed throughout the dissertation, and that recur throughout knowledge-intensive NLP.

2.1.1 Wikipedia: Encyclopedic Knowledge for Concepts and Entities

Having introduced knowledge resources at a high level, we now focus on Wikipedia as the predominant encyclopedic substrate for grounding concepts and entities. Wikipedia is a collaboratively curated online encyclopedia with multiple language editions, where knowledge is organized into interlinked articles that are continuously created, revised, and expanded by human editors. Each article is centered around a single topic and typically provides a title (a canonical name), a lead section summarizing salient information, and a hyperlink graph that connects mentions in the text to other Wikipedia pages. Wikipedia pages often include images with captions and surrounding textual context, as well as category assignments that loosely group pages by topic. From an NLP perspective, Wikipedia is

particularly valuable because it provides both (i) unstructured descriptions written in natural language and (ii) a large-scale, human-curated linking structure that can be exploited as indirect supervision. Because links are added and maintained by human editors, this dense hyperlink graph is comparatively high-quality and often connects mentions to the intended target pages. It is also widely distributed as periodic dumps, making it straightforward to build reproducible datasets tied to specific temporal snapshots.

Importantly, Wikipedia serves as the *de facto* reference corpus for many knowledge-intensive tasks, including Entity Linking (EL), retrieval-based question answering (QA), and document-grounded generation (Petroni et al., 2021; Wu et al., 2020; De Cao et al., 2021; Lewis et al., 2020b; Karpukhin et al., 2020; Kwiatkowski et al., 2019; Yang et al., 2018; Liu et al., 2018; Thorne et al., 2018). Its coverage spans both *entities* (e.g., people, locations, organizations, events) and *concepts* (e.g., common nouns and abstract notions that have encyclopedic pages), which is important for this thesis because the boundary between concepts and entities is often blurred in practical resources. In particular, Wikipedia provides a shared set of identifiers via unique page titles that enables consistent supervision and evaluation across tasks, as exemplified by KILT (Petroni et al., 2021), and supports multimodal alignment, since images and captions can be aligned to the page topic and to hyperlinked mentions in the surrounding text, enabling multimodal grounding and entity-centric supervision.

Wikipedia-grounded datasets. Additionally, Wikipedia has repeatedly served as a substrate for building high-quality datasets for both multimodal and text-only NLP tasks. On the multimodal side, Srinivasan et al. (2021) introduced Wikipedia-based Image Text (WIT), a multilingual collection of image-text pairs extracted from Wikipedia, leveraging captions and contextual signals around images across many languages; however, WIT primarily targets image-text association and does not provide the kind of explicit concept-level semantic annotations. On the text-only side, a large body of work exploits Wikipedia’s hyperlink structure to automatically derive labeled data, where hyperlinks provide naturally occurring annotations that link spans to pages and can be further typed or filtered with additional heuristics. For example, Nothman et al. (2013) introduced WikiNER, using the text and structure of Wikipedia to construct multilingual NER training data, and leveraged redirect-based

heuristics to extend mention coverage beyond explicitly hyperlinked spans. Building on related principles, Pan et al. (2017) proposed WikiANN, a language-independent framework that first assigns entity types to entries in English Wikipedia and then propagates annotations to many other languages via cross-lingual transfer, producing very large multilingual corpora but with increased noise and a focus on common coarse entity types.

Furthermore, other approaches incorporate external alignments between Wikipedia and large knowledge bases. For example, Tsai et al. (2016) used the association between Freebase and English Wikipedia as a training feature for cross-lingual NER, while Al-Rfou et al. (2015) mapped Freebase attributes to entity types to infer NER labels and additionally addressed missing annotations of non-anchored mentions through surface string-matching heuristics, combined with resampling strategies to mitigate class imbalance. More recently, Tedeschi et al. (2021) introduced WikiNEuRal, which uses Wikipedia-derived automatically generated (“silver”) annotations together with neural modeling to increase annotation density and improve coverage beyond hyperlinks, yielding higher-quality multilingual NER data by automatically identifying additional entity mentions that are not explicitly hyperlinked in Wikipedia text. Finally, Wikipedia-based hyperlink supervision has also been used successfully for large-scale dataset creation beyond NER, covering tasks such as word sense disambiguation, semantic role labeling, semantic parsing, and relation extraction, as surveyed and unified in Conia et al. (2024), as well as scalable retrieval-based EL (Wu et al., 2020).

Overall, Wikipedia’s combination of broad coverage, multilingual reach, and dense linking structure makes it a foundational resource for the datasets, benchmarks, and grounding pipelines discussed throughout this thesis. In particular, it provides the identifier space on which this thesis’s multimodal evaluation, for both concepts and entities, is ultimately grounded.

2.1.2 WordNet: Lexical Taxonomies for Concepts

Having discussed encyclopedic knowledge via Wikipedia, we now turn to WordNet as a lexical-semantic taxonomy for conceptual abstraction. WordNet is a widely used lexical database for English that organizes word meanings into a structured inventory and links them through a rich set of semantic and lexical relations (Miller, 1992). Its central organizational

unit is the *synset* (synonym set): a set of near-synonymous lemmas that express a shared sense, accompanied by a short *gloss* (definition). For instance, the word *bat* corresponds to multiple synsets, such as a flying mammal versus a piece of sports equipment, reflecting WordNet’s explicit modeling of polysemy.

A key contribution of WordNet is that it provides an explicit *taxonomy* for noun concepts, in which more specific concepts are connected to more general ones through the *hypernymy/hyponymy* relation (“is-a”): a *hyponym* is a more specific concept, and its *hypernym* is a more general parent. For example, *German shepherd* is a kind of *dog*, *dog* is a kind of *canine*, and *canine* is a kind of *carnivore*. This hierarchical structure is important in knowledge-intensive modeling because it provides a principled notion of semantic granularity: systems can be evaluated or reasoned about at the level of fine-grained concepts (e.g., *German shepherd*) or at coarser abstractions (e.g., *dog*).

Beyond hypernymy, WordNet encodes additional relations that capture different facets of lexical and conceptual structure. For nouns, *meronymy/holonymy* relations capture part-whole structure (e.g., *wheel* is part of *car*; *car* has-part *wheel*). For adjectives, relations such as *antonymy* and *similar-to* organize oppositions and fine-grained similarity neighborhoods (e.g., *hot* vs. *cold*). For verbs, WordNet includes *troponymy*, a manner-specific specialization of an action (e.g., *stroll* as a manner of *walk*), and *entailment*, where one event necessarily implies another (e.g., *snore* entails *sleep*). Collectively, these relations make WordNet more than a flat sense inventory: it is a structured semantic network that supports both linguistic interpretation and concept-level abstraction.

In multimodal research, WordNet is most prominently used as the conceptual backbone of large-scale visual taxonomies, most notably ImageNet, where class definitions are tied to synsets and inherit their granularity and organization. The taxonomical structure also enables principled label propagation and annotation expansion. For instance, returning to our earlier example, if an image is annotated with a fine-grained synset such as *German shepherd*, WordNet’s hypernym chain enables us to propagate that annotation to coarser labels (e.g., *dog*, *canine*, *carnivore*). This kind of relation-based expansion is a natural mechanism for increasing supervision density and for analyzing model behavior across levels of semantic abstraction. In this thesis, WordNet serves both as a reference point for understanding the design assumptions behind dominant vision benchmarks and as the conceptual backbone for

reasoning about semantic granularity and coverage beyond dataset-specific label sets.

2.1.3 ImageNet: Visual Taxonomies Grounded in WordNet

Having introduced WordNet’s lexical taxonomy, we now focus on ImageNet as the dominant large-scale visual benchmark that operationalizes WordNet synsets as recognition labels. Concretely, ImageNet’s classes are drawn from WordNet noun synsets that are considered visually representable, emphasizing concrete objects (e.g., animals, artifacts, foods, natural objects) while largely excluding abstract concepts and most verbs/adjectives. This yields a label space spanning multiple levels of granularity, from coarse categories to fine-grained species and product types, with a naturally long-tailed distribution shaped by both WordNet’s inventory and the availability of verifiable images.

ImageNet (Deng et al., 2009) is a large-scale visual dataset created to advance object recognition by providing labeled images for a broad set of visual concepts organized according to the WordNet lexical database. The dataset was constructed by filtering WordNet’s 82,115 noun synsets to those that are countable and visually depictable, resulting in 21,841 synset categories illustrated by over 14 million human-verified images, with an average target of roughly 1,000 images per synset. Rather than defining categories as an ad-hoc list of labels, ImageNet adopts WordNet *synsets* (sets of near-synonymous word senses) as category identifiers, grounding visual classes in an explicit lexical-semantic taxonomy. This design makes ImageNet *hierarchy-aware*: synsets are related through WordNet semantic relations, most notably hypernymy/hyponymy, enabling evaluation and analysis at multiple levels of semantic specificity (e.g., *German shepherd* $\rightarrow$ *dog* $\rightarrow$ *canine* $\rightarrow$ *carnivore*).

The motivation behind ImageNet was to create a high-scale image database aligned with a structured concept inventory, addressing a core bottleneck of early computer vision: the lack of sufficiently large and semantically organized labeled data for learning high-capacity models. By coupling web-scale image collection with a taxonomy of concepts, ImageNet enabled the community to study recognition not only at the level of coarse object categories, but also across fine-grained distinctions and long-tail concepts that arise naturally in WordNet-style inventories. The resulting dataset became a cornerstone of modern representation learning, providing both a standardized training resource and a widely adopted evaluation target that catalyzed rapid methodological progress.

Construction and annotation at scale. ImageNet was constructed with the objective of collecting labeled images to build a structured image database that covers WordNet’s visually grounded noun synsets. Its pipeline combines automated web-scale collection with extensive human verification. In broad terms, candidate images are retrieved using search queries derived from synset names and query expansion, after which annotators validate whether each image contains an instance of the target synset. This human verification step is crucial for controlling label noise at scale and for maintaining alignment between visual instances and the intended WordNet sense. To support localization and detection-style tasks, subsets of ImageNet have also been annotated with bounding boxes, enabling evaluation beyond image-level classification. Rather than simply providing a single “leaf” label per image, the ImageNet creators leveraged WordNet’s tree-like hypernym hierarchy to support automatic multi-level labeling: for an image labeled as *swan*, the corresponding ancestor concepts in WordNet imply additional valid labels such as *animal*, *vertebrate*, *bird*, and *aquatic bird*.

A key design consequence of grounding categories in WordNet is that ImageNet inherits WordNet’s sense granularity: visually similar labels may correspond to distinct synsets when they encode different senses, while visually broad concepts may be represented by high-level synsets that subsume many visually diverse subtypes. In practice, ImageNet provides images for a large portion of the WordNet noun inventory that is amenable to visual depiction, but it does not aim to be an exhaustive realization of all WordNet synsets; many synsets are abstract, non-visual, or otherwise difficult to illustrate consistently with images, and coverage is shaped by both visual depictability and data availability (Russakovsky et al., 2015). This observation suggests that while taxonomy grounding provides structure, it does not by itself guarantee broad semantic coverage of the real world.

ILSVRC and the role of standardized evaluation. A widely adopted subset of ImageNet is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC), which introduced standardized evaluation protocols for large-scale visual recognition and helped define a common benchmark for measuring progress over time (Russakovsky et al., 2015). ILSVRC is most commonly associated with the 1,000-class classification setting (ImageNet-1K), which comprises 1,281,167 training images, 50,000 validation images, and 100,000 test images.

Since the test set labels have never been publicly released, the 50,000-image validation set has become the de facto standard evaluation set used by the research community to report and compare classification results. All 1,000 categories are leaf-level synsets, meaning they have no further subdivisions in the ImageNet hierarchy, and they are predominantly drawn from a few broad semantic areas: animals (including 90 dog breeds), everyday objects (e.g., vehicles, furniture, instruments), and foods (e.g., pasta, fruits), with notably sparse coverage of other domains such as people, for which only three categories are included (*scuba diver*, *bridegroom*, and *baseball player*). We provide representative examples of ImageNet-1K categories in Figure 2.1. This skewed distribution is relevant to the arguments made in this thesis, as it shows that even a widely adopted “large-scale” benchmark can leave substantial portions of real-world semantics unrepresented. ILSVRC also included detection and localization tracks in various editions, with consistent dataset splits and evaluation metrics that enabled direct comparison between systems. The combination of scale, consistency, and public leaderboards made ILSVRC a central driver of progress in deep vision models, and ImageNet pre-training became a widely used initialization for downstream computer vision and vision-language tasks.



Figure 2.1. Examples sampled from ImageNet-1K, showing each image alongside its class label and corresponding WordNet synset.

ImageNet-21K as a large-scale pre-training resource. Beyond ImageNet-1K, a much larger portion of the ImageNet label space is commonly referred to as ImageNet-21K (or full ImageNet), which contains the full set of 21,841 synset categories and over 14 million images and is widely used for large-scale supervised pre-training (Ridnik et al., 2021). Relative to ImageNet-1K, ImageNet-21K is less of a single standardized evaluation target and more of a *training* resource that increases semantic breadth while introducing a stronger long-tail distribution over categories. From a construction perspective, it follows the same overall paradigm described above: web-scale candidate retrieval driven by synset names and query expansion, followed by human verification at scale. The key difference lies in scope, as this process is applied to a substantially larger portion of the synset inventory. We revisit the relationship between ImageNet-21K and ImageNet-1K in Section 2.2.1, where we discuss how the benchmark subset helped catalyze modern deep visual recognition.

ImageNet variants. ImageNet’s influence also revealed an important lesson: strong performance on a single, fixed benchmark distribution can mask brittleness under natural shifts, alternative data collection procedures, or changes in visual style and context. This led to a family of ImageNet-derived test sets designed to stress different failure modes. First, ImageNetV2 (Recht et al., 2019) was created to measure how well ImageNet-trained models generalize when the test set is recollected under a similar protocol but at a later time. It keeps the same label space as ImageNet-1K while changing the specific images, enabling a controlled evaluation of distribution shift that can occur even when the data collection pipeline is nominally “the same”. In a different direction, ImageNet-C (Hendrycks and Dietterich, 2019) evaluates robustness to common corruptions (e.g., noise, blur, weather-like artifacts) applied to ImageNet images, measuring the stability of predictions under systematic, non-adversarial perturbations that approximate real-world degradations. Similarly focused on challenging inputs, ImageNet-A (Hendrycks et al., 2021b) targets *natural adversarial examples*: naturally occurring images that are challenging for standard classifiers despite being correctly labeled by humans, highlighting that some failures reflect semantic or contextual brittleness rather than synthetic perturbations. Finally, ImageNet-R (Hendrycks et al., 2021a) targets *renditions* of ImageNet classes (e.g., drawings, paintings, sculptures), probing out-of-distribution generalization across visual styles while holding the underlying

concept constant.

Collectively, these successors underscore a core theme of this thesis: benchmark-centric evaluation on narrow or fixed distributions can overstate real-world competence. ImageNet’s WordNet grounding provided structure and scalability, but the extensive follow-up literature shows that both *coverage* (what concepts are represented) and *robust generalization* (how models behave under shifts in depiction, context, and distribution) require dedicated resources and evaluation protocols beyond canonical benchmarks. Accordingly, we use ImageNet as a reference point for how taxonomy-grounded benchmarks shape what models learn, while arguing that broader and more robust coverage is required for knowledge-intensive multimodal systems.

Ultimately, ImageNet has been extraordinarily influential as a large-scale visual taxonomy, and its tight alignment with WordNet provides a high-quality sense inventory and hierarchy that enable principled abstraction and evaluation across semantic granularities. However, this dependence also constrains what ImageNet can represent and measure: its label space largely mirrors WordNet’s English-centric lexical senses, emphasizing common-noun categories while offering limited coverage of named entities, multilingual realizations, and rapidly emerging concepts. As recognition and retrieval systems increasingly operate in open-world settings where entities are encyclopedic, multilingual, and continuously evolving, resources grounded in broader knowledge bases become necessary. We therefore move beyond ImageNet and WordNet by leveraging encyclopedic and multilingual resources such as Wikipedia and BabelNet, which provide wider coverage of real-world entities, cross-lingual linking, and a more faithful substrate for knowledge-intensive multimodal evaluation.

2.1.4 BabelNet: Unifying Lexicographic and Encyclopedic Knowledge

BabelNet (Navigli and Ponzetto, 2010) is a wide-coverage multilingual semantic network that unifies *lexicographic* knowledge (WordNet-like senses and semantic relations) with *encyclopedic* knowledge (Wikipedia pages), and assigns them a single shared inventory of identifiers for both concepts and named entities (Navigli and Ponzetto, 2012; Navigli et al., 2021). Its core data structure is a graph in which each node is a *Babel synset* (often referred to simply as a synset), intended to represent one meaning, concept, or entity,

and edges encode semantic relations between synsets. Crucially, BabelNet is designed to connect heterogeneous resources in a *single identifier space*: whenever a WordNet synset and a Wikipedia page are judged to refer to the same underlying concept or entity, they are merged into one Babel synset; when only one side exists (e.g., a WordNet concept without a Wikipedia page, or a Wikipedia entity without a WordNet sense), BabelNet still creates a synset anchored to the available resource.

BabelNet is constructed by merging WordNet and Wikipedia and then expanding coverage by attaching lexicalizations across many languages, leveraging the multilingual nature of Wikipedia and additional lexical resources. As a result, a single Babel synset typically contains (i) one or more canonical lexicalizations in multiple languages (surface forms), (ii) links to the originating resources (e.g., a WordNet synset offset and/or one or more Wikipedia pages), and (iii) a set of semantic relations to other synsets. This design supports cross-lingual normalization: different language-specific mentions can be mapped to the same synset identifier, enabling consistent treatment of concepts and entities across languages.

Analogously to WordNet, BabelNet provides a relational graph where synsets are connected by semantic relations (e.g., taxonomic relations such as hypernymy and hyponymy, as well as additional semantic links derived from the integrated resources). A key consequence of unifying Wikipedia and WordNet within one synset inventory is that WordNet’s high-quality conceptual structure can be propagated to encyclopedic content. For example, if the Wikipedia page for *German Shepherd* is merged with (or linked through) the corresponding synset, then the WordNet taxonomic chain can be used to connect that page to broader concepts such as *dog*, *canine*, and *carnivore* via hypernymy relations. In other words, BabelNet provides a principled mechanism for mapping encyclopedic entries onto a lexical taxonomy, enabling a single graph query interface over concept-level and entity-level knowledge.

BabelNet is particularly relevant to the contributions of this thesis because it exemplifies how concept-level and entity-level knowledge can be represented under a shared identifier space, rather than treated as disjoint inventories. This matters in multimodal resource construction and analysis: multimodal instances attached to encyclopedic pages (e.g., images appearing in a *German Shepherd* Wikipedia article) can be connected, through

BabelNet synsets, to the corresponding underlying concepts and to their taxonomic neighborhood (e.g., *dog*), supporting consistent aggregation and supervision across granularities. More generally, BabelNet illustrates a practical route to unify lexical taxonomies and encyclopedic inventories, which is aligned with the thesis objective of broad semantic coverage and principled grounding across concepts and entities.

2.1.5 CNER: Coarse-Grained Semantic Categories for Concepts and Entities

While BabelNet provides a unified identifier space for concepts and entities, resource analysis and dataset diagnostics often require a *common* semantic typing scheme that applies uniformly across both. Martinelli et al. (2024) introduce *CNER* (Concept and Named Entity Recognition), which defines a fixed inventory of 29 coarse-grained semantic categories and provides a mapping from BabelNet synsets to these categories. By assigning each synset a category in this shared label space, CNER enables systematic comparisons of semantic diversity across resources that are grounded to BabelNet (or aligned to it).

The synset-to-category mapping is created by starting from the predefined 29-category set and automatically typing BabelNet synsets using synset-level signals available in BabelNet (including metadata and links to integrated resources such as WordNet and Wikipedia), followed by refinement using additional lexical and graph-structural cues and targeted manual validation for ambiguous or low-confidence cases. This design yields an interpretable typing layer that is shared across concepts and entities.

CNER for measuring semantic coverage. Since our datasets are grounded in (or can be aligned with) BabelNet synsets, we use the CNER mapping to summarize them as distributions over the same 29 semantic categories. These categories span both common concepts and entities; for example, *ANIMAL* includes examples such as *lion* and *carnivore*, and *LOC* includes examples such as *Rome* and *Mississippi River*. We provide the complete list of categories, their descriptions, and representative examples in Table 2.1. This provides a principled way to quantify semantic coverage and compare the diversity of our resources. For instance, it enables identifying which semantic areas are well represented versus underrepresented, and contrasting concept-heavy and entity-heavy portions under a shared categorization.

Summary. In short, Wikipedia provides broad encyclopedic coverage and a shared inventory for entity-centric tasks; WordNet provides a lexical taxonomy for conceptual abstraction, ImageNet operationalizes WordNet-grounded visual categories at scale, BabelNet enables us to link lexical and encyclopedic knowledge across languages, and CNER provides a shared coarse-grained category mapping over BabelNet that we use to quantify the semantic coverage and diversity of our datasets. These resources underpin the assumptions and pipelines behind the multimodal coverage resources, entity-level datasets, and emerging-knowledge frameworks introduced in the next sections of the background chapter.

2.2 Knowledge-Intensive Multimodal Tasks

Having introduced the knowledge resources that provide the identifier spaces and grounding substrates (Section 2.1), we now turn to the three closely related multimodal tasks that leverage these resources in the training and evaluation pipelines of this thesis. We begin with *Visual Concept Recognition*, which assesses whether models can connect visual inputs to general semantic categories. We then move to *Multimodal Entity Linking*, where the objective is to resolve mentions in a multimodal context to uniquely identifiable real-world entities grounded in a knowledge base such as Wikipedia. Finally, we consider *emerging entities* in *Entity Linking*, a setting that exposes the limitations of closed-world assumptions: when the underlying knowledge base and entity inventory are treated as fixed, they inevitably become outdated as the world evolves.

2.2.1 Visual Concept Recognition

The ability to map visual inputs to general semantic categories is a prerequisite for multimodal understanding and reasoning. In practice, Visual Concept Recognition (VCR) has long served as a standard testbed for evaluating visual grounding in Vision-language Models (VLMs), making it a natural starting point for assessing systems intended for knowledge-intensive settings. The semantic coverage and representation-learning challenges exposed by this task directly motivate the questions addressed in Chapter 3.

Task Definition. VCR studies whether a model can associate visual inputs (optionally paired with text) with an underlying concept space, typically expressed through cate-

gory labels, attributes, or natural-language descriptors. Classic formulations emphasize closed-set recognition, where the label space is fixed during training and evaluation, as in ImageNet-style classification (Deng et al., 2009). More recent work increasingly targets open-vocabulary recognition, where models must align images with arbitrary textual labels or descriptions drawn from a much larger space, often enabled by contrastive pre-training on image-text pairs (Radford et al., 2021; Gu et al., 2021; Zhou et al., 2022).

Resources and Benchmarks. Progress in concept recognition has historically been driven by large-scale labeled datasets and standardized evaluation challenges. ImageNet and its WordNet-grounded hierarchy provide a canonical example of this paradigm (see Section 2.1.3), serving both as a foundational *training* resource and as a widely adopted benchmark suite for measuring recognition performance. AlexNet’s 2012 ILSVRC result demonstrated that large-scale data combined with GPU-trained convolutional networks could dramatically outperform prior approaches, and is widely viewed as a turning point that catalyzed modern deep visual recognition (Krizhevsky et al., 2012). More recent open-vocabulary and vision-language research further relies on large-scale image-text corpora, including Conceptual Captions (Sharma et al., 2018), which provides millions of web-mined image-caption pairs for learning language-aligned visual representations, and web-scale datasets such as LAION (Schuhmann et al., 2022), which scale this supervision to billions of image-text pairs.

For *evaluation*, the ILSVRC defined a curated 1,000-class subset, ImageNet-1K, which remains the dominant closed-set benchmark. In this thesis, we use ImageNet-1K as a reference point because it has anchored much of the field’s benchmarking efforts, but we argue that success on ImageNet-1K alone can mask deficiencies in semantic breadth and out-of-distribution robustness that are central to knowledge-intensive multimodal systems. Several benchmarks have since been introduced to probe dimensions that ImageNet-1K does not adequately cover. ImageNet-CoG (Sariyildiz et al., 2021) was specifically designed to evaluate generalization to concepts beyond those in ImageNet-1K, by defining progressively more distant concept sets within ImageNet-21K; however, in our analysis we observe that it still draws heavily from similar high-level semantic regions (e.g., animals and other concrete object categories) rather than substantially shifting the overall topical balance of what is

being evaluated.

Other benchmarks target specific domain or distributional gaps. The *iNaturalist* (Horn et al., 2018) benchmark, often referred to as iNat, was introduced to better reflect real-world recognition in biodiversity domains, where category frequencies are inherently *long-tailed* and fine-grained distinctions are common (e.g., visually similar species under large variations in pose, illumination, seasonality, and capture conditions). Compared to web-curated datasets, iNat emphasizes rare-category performance and robustness under natural distribution shifts. The benchmark has been released in multiple editions with increasing scale: the original dataset covers >5,000 species and 859K images (Horn et al., 2018), while later challenge editions commonly used in modern evaluation include iNat2018 with 8,142 species and iNat2021 with 10,000 species. Classes are species-level categories spanning broad taxonomic groups such as birds, insects, plants, fungi, reptiles, and mammals, making fine-grained recognition and long-tail generalization central.

Addressing a complementary limitation, *ObjectNet* (Barbu et al., 2019) was designed as a bias-controlled testbed to test object recognition robustness beyond ImageNet-style imagery. A key motivation is that ImageNet models can exploit spurious correlations between objects and typical backgrounds or contexts; ObjectNet therefore uses a controlled collection protocol to introduce systematic shifts in viewpoint, rotation, and background, aiming to isolate object understanding from dataset-specific shortcuts. The benchmark provides a 50,000-image test set covering 313 object categories, with 113 categories overlapping ImageNet, and is intentionally released without a standard training set to discourage benchmark-specific overfitting. In practice, ObjectNet is often used as an out-of-distribution robustness evaluation: models that perform strongly on ImageNet can exhibit substantial performance drops on ObjectNet under these controlled shifts.

Beyond these individual benchmarks, models are often assessed using evaluation suites that span multiple visual domains, helping probe robustness beyond natural-image object recognition. For instance, FER2013 (Goodfellow et al., 2015) targets *facial expression recognition* with 7 discrete emotion classes from 48×48 grayscale face crops, making performance sensitive to pose, illumination, and subtle texture cues. STL-10 (Coates et al., 2011) evaluates *object recognition* over 10 classes with relatively few labeled examples and additional unlabeled data, emphasizing representation learning under limited supervision.

EuroSAT (Helber et al., 2019) tests *land-use and land-cover classification* from Sentinel-2 satellite imagery over 10 scene categories (e.g., residential, industrial, forest), introducing a remote-sensing domain shift with different spatial statistics and acquisition conditions. Similarly, NWPU-RESISC45 (Cheng et al., 2017) provides *aerial scene classification* at a larger label granularity, covering 45 remote-sensing scene classes with substantial intra-class variability across geography, resolution, and imaging conditions. GTSRB (Stallkamp et al., 2011) focuses on *fine-grained traffic sign recognition* with 43 categories under real driving variability (viewpoint, motion blur, and illumination), while PatchCamelyon (Veeling et al., 2018) evaluates *histopathology patch classification* (tumor versus non-tumor), probing sensitivity to subtle, domain-specific visual evidence in medical imagery. Despite their diversity, these benchmarks still capture only selected slices of real-world semantics and often rely on coarse or domain-specific label spaces, underscoring that strong performance across such suites does not necessarily translate into broad concept coverage or systematic generalization to long-tail and underrepresented semantic categories.

Finally, complementary to these predominantly concrete benchmarks, *BabelPic* (Calabrese et al., 2020b) is a dataset that targets *non-concrete* concepts that are difficult to depict directly and often require contextual, metaphorical, or compositional visual cues. BabelPic is explicitly designed to probe whether multimodal models can align images and text to abstract meanings beyond object-centric recognition, providing a useful counterpoint to standard concept recognition benchmarks that largely emphasize visually salient, concrete categories.

Current Approaches. A dominant paradigm in modern recognition is to learn *shared image-text representation spaces* that enable zero-shot and few-shot transfer via natural-language prompts. CLIP (Radford et al., 2021) popularized large-scale contrastive pre-training on noisy web image-text pairs, showing that matching images to captions yields representations that can be *queried with textual class names* for strong zero-shot recognition. Building on this recipe, SigLIP (Zhai et al., 2023) introduced a *pairwise sigmoid loss* for language-image pre-training that removes softmax normalization over the batch, improving scaling behavior and performance across batch sizes while retaining the same prompt-based zero-shot interface. Recent work has also targeted *efficiency bottlenecks* in multimodal

recognition: Vasu et al. (2025) propose FastVLM, an efficient vision-encoding design that reduces the number of visual tokens (and thus latency/compute) while preserving high-resolution visual fidelity, making prompt-based recognition more practical in VLM pipelines. Complementarily, Fini et al. (2025) study large multimodal models as *open-world image classifiers*, formalizing evaluation protocols and metrics for free-form predictions when the label set is not predefined, thereby pushing beyond closed-set accuracy toward semantic alignment in natural-language outputs.

Large multimodal foundation models extend this paradigm with stronger cross-modal fusion and instruction-following, enabling more flexible concept querying via natural language. Flamingo (Alayrac et al., 2022) showed that bridging strong pretrained vision and language backbones with dedicated cross-attention mechanisms yields *in-context few-shot* adaptation across diverse multimodal tasks from exemplars in the prompt. BLIP-2 (Li et al., 2023) introduced a compute-efficient strategy that keeps both the vision encoder and LLM frozen and learns a lightweight *Querying Transformer (Q-Former)* to align modalities, substantially reducing trainable parameters while enabling strong zero-shot generation and instruction following. PaLI (Chen et al., 2023b) advanced the generative formulation by jointly scaling language and vision and using a unified text generation interface for many multimodal tasks, including multilingual settings, further blurring the line between recognition and generation. Instruction-tuned VLMs such as InstructBLIP (Dai et al., 2023a) and LLaVA (Liu et al., 2023a) further demonstrated the effectiveness of adapting vision-language models to instruction-following settings, supporting open-ended multimodal dialogue, reasoning, and task generalization. Overall, these models increasingly bridge recognition and generation by producing free-form outputs that can express fine-grained or compositional concepts not explicitly enumerated as labels.

More recent general-purpose multimodal models further extend this trend by treating multimodal understanding as an open-ended instruction-following problem. Systems such as Gemini (Team et al., 2024) and Qwen2.5-VL (Bai et al., 2025) can accept images together with natural-language prompts and produce textual outputs such as concept names, entity descriptions, rationales, and answers to visual questions. This makes them increasingly relevant to knowledge-intensive multimodal tasks, where the desired output often requires both visual grounding and access to external or parametric knowledge. However, their flexibility

also introduces important challenges: generated outputs may be difficult to normalize to controlled concept or entity inventories, may hallucinate plausible but unsupported information, and may rely on outdated parametric knowledge when entities or facts change over time. A complementary line of work has focused on specific semantic coverage gaps that remain difficult for both discriminative and generative VLMs. In particular, *non-concrete* concept understanding represents an important challenge because such concepts are not directly depicted by isolated objects and often require contextual, metaphorical, or compositional visual cues. For instance, EViLBERT (Calabrese et al., 2020a) builds on the motivation of BabelPic to study multimodal representations of abstract concepts, introducing an evaluation setting that probes whether models can go beyond object-centric cues when grounding concept meaning.

Coverage and Generalization Challenges. Despite strong headline performance, widely used benchmarks can exhibit narrow semantic scope and dataset-specific regularities that overestimate robustness under distribution shift (Torralba and Efros, 2011; Geirhos et al., 2020; Taori et al., 2020). This issue is particularly evident in multimodal settings, where shifts in visual style, context, or label granularity can lead to substantial degradation even for high-performing models. These limitations motivate the need for resources and evaluation protocols that broaden semantic coverage and explicitly emphasize generalization beyond conventional benchmark distributions.

Table 2.2 summarizes the semantic coverage of widely used benchmarks by reporting (i) the absolute number of concepts (i.e., dataset classes) and (ii) the CNER semantic categories that these concepts collectively cover. For each dataset, we assign CNER categories to its classes; when a dataset’s provided labels are not directly grounded in standard knowledge resources (e.g., Wikipedia, WordNet, or BabelNet), we perform this assignment manually to ensure consistent normalization. We also include ImageNet-1K as a canonical large-scale taxonomy-derived benchmark, but omit closely related ImageNet variants (e.g., ImageNet-v2, ImageNet-R) because many share the same label set; even ImageNet-CoG was shown to largely introduce different concepts within the same underlying CNER semantic categories. As seen in Table 2.2, even datasets with large label spaces can remain limited in semantic category coverage; moreover, even when combining all listed benchmarks we reach only 18

of the 29 CNER categories, motivating the need for wider-coverage benchmarks, which we discuss in Chapter 3.

2.2.2 Multimodal Entity Linking

While visual concept recognition maps inputs to broad semantic categories (Section 2.2.1), many knowledge-intensive applications require identifying specific real-world entities from combined visual and textual evidence, a setting commonly referred to as Multimodal Entity Linking (MEL). This introduces an additional disambiguation challenge, since visually or textually similar mentions may refer to different entities in a knowledge base. In this section, we lay the groundwork for this task and for the generalization challenges it faces, which we address in Chapter 4.

Task Definition. Multimodal Entity Linking (MEL) extends text-only entity linking by resolving an entity mention using both textual context and associated visual evidence, linking the mention to a canonical entry in a knowledge base such as Wikipedia. Formally, given a mention and its multimodal context, the system predicts an entity identifier from an inventory of candidates, typically using a scoring function between the context representation and an entity representation. Because entities are uniquely referential, MEL requires both recognition (leveraging visual cues) and disambiguation (leveraging contextual and encyclopedic information).

Resources and Benchmarks. Early MEL datasets investigated disambiguation in multimodal user-generated content, emphasizing how images can resolve ambiguous mentions in social media and short documents (e.g., DZMNED) (Moon et al., 2018). Subsequent work introduced additional MEL datasets and settings beyond social media, including Wikipedia-centric and domain-specific collections, reflecting the breadth of multimodal linking conditions studied in the literature (e.g., RichpediaMEL and MELArt) (Hu et al., 2025; Sierra-Múnera et al., 2024). More recent Wikipedia-grounded benchmarks such as WikiDiverse (Wang et al., 2022c) and WikiMEL (Yang et al., 2023) provide standardized evaluation settings for MEL and expose dataset shifts across entity distributions, topical mixtures, and multimodal context. These benchmarks are central to this thesis because

they instantiate MEL under encyclopedic grounding while still highlighting generalization failures across datasets.

Current Approaches. Most MEL systems follow the same high-level structure as text-only EL: encode the mention context, encode candidate entities, and match them through retrieval and/or reranking. A common scalable formulation is bi-encoder retrieval, popularized in EL by BLINK (Wu et al., 2020), which maps contexts and entities into a shared embedding space for efficient nearest-neighbor search. MEL extends this template by integrating visual evidence into the context encoder and, in some cases, into entity representations, using fusion mechanisms that range from early concatenation to cross-attention-based multimodal interaction.

Model families and design choices. At a high level, existing MEL methods can be grouped by how they (i) retrieve candidates and (ii) fuse modalities for disambiguation. In a typical two-stage pipeline, a first-stage retriever proposes a small set of entity candidates (e.g., via bi-encoder similarity), and a second-stage reranker performs more expensive cross-attention over the multimodal context and candidate descriptions to select the final entity. This mirrors the retriever-reranker recipe in text-only EL, but the reranker must additionally decide when the image is informative versus when it is misleading or irrelevant to the mention.

A key design choice is the *fusion mechanism* used to integrate visual features and text. Early-fusion approaches concatenate or jointly encode the mention context with visual embeddings so that the encoder produces a single multimodal representation for retrieval. Late-fusion approaches compute text-only and image-only similarity signals to candidates and then combine them via weighting or learned gating, which can be more robust when one modality is noisy or missing. Cross-attention based interaction (e.g., multimodal Transformers) provides the most expressive fusion, but is typically reserved for reranking due to computational cost.

Another axis is how *entity representations* are constructed. In Wikipedia-grounded settings, candidates are often represented using their title and a short textual description (e.g., the first sentence, the lead section, or a definition-like snippet), sometimes augmented with category/type information or an associated representative image. Entity representations

can also be made *multi-vector* (e.g., separate fields for title/description/image) to support field-aware retrieval and to reduce over-reliance on a single text field that may be missing or biased.

VLM-based and generative formulations. With the rise of vision-language foundation models, a recent trend is to treat MEL as a *multimodal question answering* or *instruction following* problem, where the model is prompted with the mention, its context, and the image, and asked to output the linked entity (e.g., as a Wikipedia title). These approaches can reduce engineering overhead, but introduce additional challenges: (i) outputs must be normalized reliably to KB identifiers, (ii) models may hallucinate plausible entities that are not in the inventory, and (iii) evaluation becomes sensitive to string matching and alias handling. As a result, even when using VLMs, practical MEL systems often retain retrieval- or index-based constraints (e.g., restricting generation to a candidate list) to preserve scalability and closed-world correctness.

Generalization and Benchmark-Centric Bias. A recurring limitation in MEL is benchmark-centric evaluation: models are typically trained and tuned on a single dataset and then assessed explicitly using that same dataset’s test set. As a result, strong in-domain performance may not translate into robustness under dataset shift or across benchmarks with different entity distributions and visual styles. This thesis emphasizes cross-dataset transfer as a primary diagnostic and motivates large-scale resources that reduce dependence on dataset-specific distributions.

2.2.3 Entity Linking with Emerging Entities

The previous sections addressed grounding at the concept level and at the entity level under the assumption that the reference knowledge base provides adequate coverage. Many standard approaches still rely on this closed-world assumption, even though real-world knowledge bases are inherently incomplete and lag behind the continuous emergence of new entities, a limitation that is especially pronounced for Entity Linking (EL). This section reviews emerging entities in EL and the challenges they introduce, which are addressed in Chapters 5 and 6.

Entity Linking. EL is the task of associating entity mentions in text with identifiers in a reference knowledge base, most commonly Wikipedia, and it is a core building block for knowledge-intensive pipelines such as retrieval-based question answering and document-grounded generation (Mihalcea and Csomai, 2007; Petroni et al., 2021). Because the same surface form can refer to multiple entities depending on context (e.g., *Apple* as a company vs. a fruit), EL typically combines mention identification with disambiguation to a single knowledge-base entry (Hoffart et al., 2011; Ganea and Hofmann, 2017; Kolitsas et al., 2018). Modern EL systems are increasingly retrieval-driven: they rely on neural encoders to retrieve and rank candidates at scale, commonly following bi-encoder retrieval with optional reranking, as exemplified by dense entity retrieval approaches (Wu et al., 2020).

Text-only EL has been driven by a small number of widely used benchmarks that differ in genre, mention density, and degree of ambiguity, but largely share a common design choice: mentions are linked to a fixed Wikipedia snapshot (or a KB derived from it), and evaluation assumes that the correct entity exists in that inventory. These benchmarks are discussed in more detail below (Section 2.2.3), but a common limitation is that they treat the underlying knowledge base as temporally static, under-specifying the fact that entity inventories and descriptions evolve and enabling temporal leakage when train and test draw from overlapping snapshots.

Datasets and Benchmark Suites. We now briefly summarize the most commonly used EL benchmarks and how they are typically used in evaluation. Early work standardized evaluation on newswire and web documents, notably AIDA-CoNLL (Hoffart et al., 2011), which remains a canonical benchmark for disambiguation under relatively clean, in-domain text. Subsequent datasets expanded coverage to noisier and more diverse settings, including TAC-KBP EL, which introduced forum and news genres with more challenging mention contexts and end-to-end evaluation protocols that distinguish mention detection, candidate generation, and disambiguation in its scoring setup. Other benchmarks emphasize short, ambiguous contexts and are commonly used as small, news-derived auxiliary test collections. The MSNBC dataset consists of a small set of news articles annotated with entity links, and is often used to evaluate disambiguation when only limited local context is available. The AQUAINT collection similarly draws from newswire documents and is frequently used

as an additional auxiliary evaluation set in cross-domain comparisons. ACE2004 provides another news-domain test collection with entity annotations, and is commonly evaluated alongside MSNBC and AQUAINT to probe robustness of disambiguation under limited context and domain shift (Ganea and Hofmann, 2017; Kolitsas et al., 2018).

In practice, many of these corpora are also used in the more constrained *entity disambiguation* (ED) setting, where gold mention spans are given and the task reduces to selecting the correct KB entry. This ED protocol is especially common for the smaller collections (including MSNBC, AQUAINT, and ACE2004), and is useful for isolating errors due to ambiguity resolution from upstream mention detection. Across benchmarks, sizes vary substantially: AIDA-CoNLL provides a relatively large, document-level evaluation set, whereas the auxiliary news collections are much smaller and are primarily used for robustness comparisons rather than for training.

In addition to these human-annotated corpora, Wikipedia itself provides large-scale weak supervision via hyperlinks: anchor spans can be treated as mention boundaries and their targets as entity identifiers. This “hyperlink-as-label” signal has been widely used to create pre-training data for entity encoders and to train scalable candidate generation models, despite known biases (e.g., only a subset of mentions are linked, links skew toward salient entities, and annotation conventions vary across pages).

As neural retrieval became central, benchmark suites increasingly focused on scalability and comparability across systems. GERBIL (Usbeck et al., 2015) provided a unified evaluation framework covering many standard EL datasets, making it easier to compare systems under consistent metrics and annotation conventions. More recently, KILT (Petroni et al., 2021) reframed EL as part of a broader family of knowledge-intensive tasks grounded in Wikipedia, with standardized train/dev/test splits and evidence requirements that facilitate end-to-end modeling and cross-task transfer. Alongside these, large-scale hyperlink-derived corpora from Wikipedia have been widely used for supervision and pre-training (often as weakly labeled training data), complementing smaller human-annotated benchmarks.

Emerging Entities and Temporal Drift. A standard evaluation assumption in EL is that the entity inventory and the knowledge base snapshot are fixed at inference time (Shen et al., 2014; Usbeck et al., 2015). While convenient for benchmark design, this fixed-inventory

setting becomes limiting in real-world environments where entities and their associated facts evolve over time, and where systems must operate under temporal change and distribution shift. Emerging entities expose the static-inventory limitation directly: when an entity is missing from the current inventory, a linking system cannot resolve it without updating the underlying resource or changing the model.

This issue is closely related to the long-standing *NIL/out-of-KB* setting popularized by TAC-KBP, where systems must decide that no suitable KB entry exists for a mention and then cluster such NIL mentions. From the perspective of this thesis, emerging entities can be interpreted as a temporal manifestation of out-of-KB linking: what is “out-of-KB” today may become in-KB tomorrow once the inventory is updated. Accordingly, emerging-entity EL provides a concrete operationalization of knowledge dynamics and motivates methods and evaluation protocols that explicitly separate training and testing snapshots to avoid temporal leakage and to measure robustness under drift.

Recent work has made this failure mode explicit by targeting temporal drift and emerging entities directly. TempEL provides time-based Wikipedia snapshots and evaluation splits to study EL under temporal change, enabling controlled experiments that train on one time period and evaluate on later snapshots, and highlighting the impact of temporal leakage and contamination when train and test are not time-separated (Zaporojets et al., 2022).

Beyond evaluation, CYCLE (Zhang et al., 2024a) proposes cross-year contrastive objectives to improve robustness across time, reflecting the need for entity representations that transfer across evolving distributions rather than overfitting to a single snapshot. Complementary directions address this issue more practically. For example, STAMO (Ran et al., 2022) attempts to learn EL features for emerging entities, while EDIN (Kassner et al., 2022) frames the problem as *unknown entity discovery and indexing*, emphasizing that practical pipelines must first detect and represent entities absent from the inventory before they can be linked. More broadly, survey work highlights that modern EL systems are increasingly retrieval-based and scalable, yet still largely evaluated under fixed candidate assumptions, motivating research that relaxes inventory constraints and improves robustness under temporal shift (De Cao et al., 2021; Ayoola et al., 2022; Sevgili et al., 2022).

Implications for Downstream Knowledge-Intensive Systems. Because EL is often an upstream component for knowledge access and attribution, failures on emerging entities propagate to downstream systems that rely on correct identification for retrieval, grounding, and factual consistency (Shen et al., 2014). This motivates approaches that can (i) acquire evidence about previously unseen entities from external sources, and (ii) update knowledge inventories so that emerging entities become usable by existing components without expensive retraining (Zaporojets et al., 2022; De Cao et al., 2021).

2.3 Dynamic Knowledge Integration

A core limitation of parametric knowledge in large pretrained models is that it is *fixed* once training is complete, even though real-world information changes continuously. A widely adopted strategy to mitigate this limitation is to *integrate external evidence at inference time*, allowing models to condition their outputs on retrieved documents rather than relying solely on internalized information. This paradigm underlies retrieval-augmented modeling, which has become central in knowledge-intensive NLP, including open-domain question answering, fact-sensitive generation, and other settings where correctness depends on access to external sources (Guu et al., 2020; Lewis et al., 2020b).

2.3.1 Knowledge Editing and Model Updating

Retrieval is not the only approach to handling changing knowledge. A complementary line of work studies *knowledge editing*: directly modifying a trained model so that it updates, inserts, or deletes specific facts without full retraining. This direction is motivated by the observation that deploying retrieval infrastructure is not always possible, and that some applications require fast, targeted updates (e.g., to correct a salient factual error).

Editing methods typically aim to (i) implement the desired local change, (ii) preserve unrelated behavior, and (iii) generalize the edit appropriately (e.g., to paraphrases) while avoiding unintended side effects. Within the scope of this thesis, these methods provide an alternative perspective on *emerging knowledge*: instead of expanding an external inventory or retrieving up-to-date evidence, one can attempt to *repair* or *refresh* parametric knowledge directly. However, because emerging entities and newly created concepts often require *new*

identifiers and *new supporting evidence*, retrieval- and inventory-based mechanisms remain especially natural for knowledge-intensive grounding tasks.

2.3.2 Retrieval-Augmented Modeling

Retrieval-augmented modeling typically decomposes prediction into two coupled components: (i) retrieving relevant evidence from a large corpus (e.g., Wikipedia), and (ii) generating or selecting an output conditioned on the retrieved evidence. Early and influential approaches instantiate retrieval using dense bi-encoders, where queries and passages are mapped into a shared embedding space to support scalable nearest-neighbor search (Karpukhin et al., 2020). Building on this paradigm, retrieval can be integrated into end-to-end generation, enabling models to incorporate evidence dynamically during decoding (Lewis et al., 2020b; Guu et al., 2020).

Retrieval for knowledge-intensive NLP. Dense retrieval frameworks such as DPR introduced a practical open-domain QA recipe based on training dual-encoder query and passage representations in a shared embedding space with efficient nearest-neighbor search, demonstrating that learned dense retrieval can substantially improve evidence access compared to sparse retrieval (Karpukhin et al., 2020). REALM further showed that retrieval can be incorporated into language modeling itself by treating retrieved documents as latent variables and pre-training a retriever-reader system so that retrieval directly supports downstream prediction, typically retrieving from a large textual corpus such as Wikipedia (Guu et al., 2020). These developments helped establish retrieval as a practical mechanism for connecting neural models to large, updatable corpora rather than treating knowledge as purely parametric.

Beyond open-domain QA, retrieval has also been adopted to support NLP tasks by restricting prediction to a small candidate set supported by retrieved evidence. In this *retrieve-then-predict* setting, retrieval serves as candidate generation over a large label space (e.g., entities or relations), and the model then performs contextual matching, reranking, or structured prediction conditioned on the input and the retrieved candidates. ReLiK (Orlando et al., 2024) exemplifies this direction for entity-centric information extraction. It retrieves candidate entities (and relations) from an index built from Wikipedia-derived entity entries

(e.g., titles and short definitions), and then matches candidates to spans (and pairs of spans) in the input text to produce mention-entity links and relations. Fusion Entity Decoding (FusionED) follows a similar approach for Entity Linking and Disambiguation, retrieving a candidate set and then performing *generative* disambiguation by decoding the index of the correct entity among the retrieved candidates, with candidates grounded in entity titles and descriptions (Wang et al., 2024).

Relatedly, retrieval augmentation has been studied for large-label text classification beyond entity-centric tasks. For instance, retrieval-augmented multi-label classification augments a Transformer encoder with cross-attention over similar training documents retrieved for the input, which can improve performance particularly for infrequent (long-tail) labels (Chalkidis and Kementchedjhieva, 2023).

Generative retrieval-augmented models. Retrieval-Augmented Generation (RAG) introduced an end-to-end formulation that couples a dense retriever with a sequence-to-sequence generator, conditioning generation on retrieved passages and (in its original formulation) marginalizing over multiple retrieved documents during decoding (Lewis et al., 2020b). In the standard open-domain QA setting, the retriever indexes a large passage collection, commonly derived from the full text of Wikipedia segmented into passages, so that generation is explicitly conditioned on retrieved evidence text rather than only on entity-level descriptors (Lewis et al., 2020b; Karpukhin et al., 2020). Fusion-in-Decoder (FiD) strengthens this paradigm for open-domain question answering by encoding multiple retrieved passages independently and fusing them in the decoder via cross-attention, improving the model’s ability to aggregate evidence across sources (Izacard and Grave, 2021). ATLAS further scales this family of approaches by jointly training the retriever and generator and demonstrating strong performance across multiple knowledge-intensive tasks within a unified retrieval-and-generation framework (Izacard et al., 2023). At a larger scale, RETRO augments language modeling with a massive external memory by retrieving relevant chunks from trillions of tokens and using chunked cross-attention, illustrating how retrieval can substantially expand accessible information beyond what is stored parametrically (Borgeaud et al., 2022).

While the above focuses on retrieval to support generation, retrieval-augmented generation is most commonly studied in open-domain QA and related knowledge-intensive set-

tings, where retrieval is typically implemented via dense passage retrieval over Wikipedia or other large corpora (e.g., DPR-style retrieval) (Karpukhin et al., 2020). Benchmarks such as KILT standardize this setting by grounding multiple knowledge-intensive tasks—including open-domain QA and fact checking—on a shared Wikipedia snapshot, enabling evaluation not only of answer quality but also of evidence provenance (Petroni et al., 2021). Nevertheless, retrieval does not guarantee factuality: generation can still introduce unsupported statements even when relevant evidence is retrieved, motivating the parallel line of work on evidence-grounded and interpretable factuality evaluation discussed later in this chapter.

Despite strong empirical gains, retrieval-augmented methods introduce new failure modes. First, performance depends critically on *evidence selection*: if retrieval misses key evidence or returns misleading passages, the model may produce incorrect outputs (Lewis et al., 2020b; Guu et al., 2020). Second, even when relevant evidence is retrieved, generative models may still produce *unsupported* statements, either by hallucinating details or by blending multiple sources incorrectly. Such failures in terms of faithfulness have been widely documented in abstractive summarization and factual generation, where fluent outputs can still contradict the source document. These observations motivate evaluation frameworks that explicitly test whether model outputs are supported by the evidence they condition on (Section 2.4).

2.3.3 Knowledge Integration under Dynamic Knowledge

In real-world settings, the underlying knowledge changes faster than static resources and model parameters can be updated. This motivates methods that *incorporate newly available information* through retrieval and controlled generation, rather than assuming a complete and fixed knowledge snapshot. In knowledge-base-centric tasks such as entity linking, the reliance on fixed inventories and static repositories creates a mismatch with real-world deployment, where new entities emerge and previously valid facts may be revised over time (Zaporojets et al., 2022). Retrieval-augmented approaches provide a natural mechanism to reduce this mismatch by enabling models to condition on recently published or updated evidence, rather than requiring costly retraining whenever the knowledge snapshot changes (Lewis et al., 2020b; Guu et al., 2020).

A recurring theme across knowledge-intensive systems is that *closed-world assump-*

tions simplify benchmarking but limit practical applicability. When models can consult external corpora at inference time, they can, in principle, support decisions and outputs using evidence that reflects the current state of information. However, this shift also places greater emphasis on robustness to noisy or incomplete evidence and, more importantly, evaluation protocols that can validate evidence support without relying on a single gold reference.

2.4 Knowledge-Intensive Evaluation

Evaluating knowledge-intensive systems requires more than measuring fluency or topical relevance. The central question is whether model outputs are consistent with the available evidence and with correct facts about the world. This becomes especially challenging in settings where knowledge changes over time or where outputs are expected to reflect newly available information, as in such cases a single fixed gold reference is often unavailable or quickly becomes outdated.

2.4.1 Limitations of Reference-Based Similarity Metrics

A common starting point for evaluating generation is surface similarity to a reference, typically using n -gram overlap measures such as ROUGE (Lin, 2004). These metrics are inexpensive and reproducible, and have historically been widely adopted in summarization and related generation settings. However, overlap is only a weak proxy for correctness: two outputs can be semantically equivalent with low overlap, and conversely, an output can obtain high overlap while containing factual errors or unsupported details. To mitigate the limitations of surface overlap, semantic metrics compare reference and hypothesis in embedding space, including BERTScore (Zhang et al., 2019) and learned evaluators such as BLEURT (Sellam et al., 2020). While these metrics better capture paraphrase-level similarity, they still do not directly assess whether an output is *supported by evidence*, and they remain fundamentally dependent on the existence and quality of reference texts.

2.4.2 Factual Consistency and Evidence-Grounded Evaluation

A large body of work has shown that high-quality text generation can still contain factual inconsistencies, omissions, and hallucinations, especially in summarization and knowledge-

grounded generation settings (Maynez et al., 2020; Kryscinski et al., 2020). As a consequence, recent research increasingly treats factuality (or factual consistency) as a central evaluation target, complementing generic similarity metrics.

A common strategy is to evaluate factual consistency by verifying whether a generated output is supported by the available evidence, typically the source document(s) on which the model is conditioned. Examples of this include FactCC (Kryscinski et al., 2020), FEQA (Durmus et al., 2020), and NLI-based frameworks such as SummaC (Laban et al., 2022). Additionally, question-generation and question-answering style evaluators have been proposed for reference-free consistency checking, including Q^2 (Honovich et al., 2021). These approaches bring evaluation closer to evidence verification, but they differ in the unit they validate (spans, answers, or propositions), the scope of evidence they assume (single document vs. multiple sources), and the extent to which they provide interpretable signals and explicit attribution.

Beyond the document-consistency methods discussed above, several complementary lines of work further illustrate the trade-offs between performance and interpretability in factuality evaluation. AlignScore (Zha et al., 2023) introduces an alignment-based evaluator trained across multiple verification tasks, including natural language inference (NLI). While it achieves strong performance for factuality evaluation, it provides only a single global score for an entire summary, which limits interpretability by making it difficult to localize which parts of the output are factual or non-factual. Moreover, AlignScore aligns the summary to relatively large text blocks from the input document, which can reduce the diagnostic value of the returned alignments. On the other hand, Dependency Arc Entailment (DAE) (Goyal and Durrett, 2020) evaluates factual consistency by leveraging entailment signals between dependency-parse structures of document sentences and the summary. However, because abstractive summaries may compress multiple facts into a single sentence, purely sentence-level structural comparisons can be suboptimal in highly abstractive regimes. Overall, despite substantial progress, there remains considerable room for improving NLI-based metrics in both performance and interpretability.

LLM-based metrics. Several studies (Shen et al., 2023; Chen et al., 2023a; Wang et al., 2023a) analyze the use of Large Language Models (LLMs) as summarization evaluators

across multiple settings. Their findings suggest that, although LLM-based evaluators can outperform many traditional automatic metrics, they are not yet reliable replacements for human evaluation. Building on this direction, TrueTeacher (Gekhman et al., 2023a) investigates whether LLMs can be used to generate large-scale factuality supervision and then trains smaller student evaluators on such data, surpassing prior NLI-based metrics. Another prominent trend relies on prompt-based evaluation (Luo et al., 2023b; Fu et al., 2023; Liu et al., 2023b; Gao et al., 2023), where an LLM is instructed using the evaluation task description together with the input document and the system-generated summary. GPTScore (Fu et al., 2023) demonstrates that such prompting can support multi-aspect, customizable, and training-free evaluation. Relatedly, G-Eval (Liu et al., 2023b) explores chain-of-thought and structured form-filling instructions, showing that when instantiated with GPT-4 it achieves strong correlation with human consistency annotations on SummEval (Fabbri et al., 2021).

2.4.3 Evaluation Under Evolving Knowledge

In many knowledge-intensive scenarios, especially when knowledge changes over time or when systems are expected to incorporate newly available information, a single gold reference is inherently unavailable. For example, an updated description of an entity or a rapidly developing event may be correct and evidence-supported even if it does not match any previously written references. This motivates evaluation protocols that verify *support in evidence* rather than similarity to a fixed target, and that provide fine-grained feedback about which parts of an output are unsupported.

In practice, these evaluation settings are often instantiated using Wikipedia-derived evidence, either by tying tasks to specific Wikipedia snapshots/dumps (to ensure reproducibility) or by explicitly checking claims against Wikipedia passages. This includes long-document summarization and lead-section generation, where systems attempt to synthesize evidence from multiple documents into Wikipedia-style text, as in WikiSum (Liu et al., 2018). Wikipedia is also used in structured generation settings such as biography generation (e.g., WikiBio; (Lebret et al., 2016)), which highlights that reference overlap can be a poor proxy for factual correctness (Stranisci et al., 2023). In addition, Wikipedia-derived benchmarks target more constrained summarization objectives, such as aspect-based

summarization from Wikipedia pages (Hayashi et al., 2021). Across these settings, evaluation is tightly coupled with factuality, and correctness must ultimately be judged with respect to evidence and world knowledge, rather than reference overlap alone. A closely related line of work comes from claim verification, where benchmarks such as FEVER (Thorne et al., 2018) evaluate whether individual claims can be supported or refuted using Wikipedia evidence. This perspective closely aligns with evidence-grounded evaluation of generation, motivating frameworks that both *localize* unsupported content and *attribute* supporting evidence for their decisions. However, these benchmarks still operate over static claim-evidence pairs and do not explicitly target the evaluation of systems that must generate or update knowledge-intensive outputs as the underlying evidence evolves.

These considerations underscore the need for evaluation methods that explicitly target the dynamic and inherently reference-free nature of real-world knowledge. In particular, beyond producing a single aggregate score, evaluators should enable factual verification against currently available evidence while providing interpretable signals (e.g., claim-level decisions and evidence attribution) that make failures diagnosable and actionable.

Category	Description	Examples
ANIMAL	Living beings (excluding humans) with the ability to move and perceive their surroundings.	<i>dog, cat, mammal</i>
ARTIFACT	Man-made objects, tools, and products.	<i>vehicle, software, mouse</i>
ASSET	Resources with economic or intrinsic value.	<i>capital, stock, wealth</i>
BIOLOGY	Biological entities such as cells or organisms.	<i>protein, cell, living organism</i>
CELESTIAL	Astronomical objects like planets, and galaxies.	<i>comet, nebulae, Sun</i>
CULTURE	Cultural aspects, traditions, or ideologies.	<i>religion, feminism, socialism</i>
DATETIME	Dates and temporal expressions.	<i>18 March, Saturday, 1979</i>
DISEASE	Medical conditions or disorders.	<i>infection, allergy, metastasis</i>
DISCIPLINE	Specific fields of study, knowledge, or expertise.	<i>discipline, sport, football</i>
EVENT	Phenomena or activities at a specific time/place.	<i>crime, temperature change</i>
FEELING	Emotions, sensations, and experiences related to human or animal consciousness.	<i>affection, attachment</i>
FOOD	Edible items and beverages.	<i>beverage, dish, pork</i>
GROUP	Group of people or animals.	<i>staff, social group, panel</i>
LANGUAGE	Language-related items, such as words, or phrases.	<i>discourse, context, lexeme</i>
LAW	Legal principles, regulations, and rules governing society and various aspects of life.	<i>law, civil law</i>
LOC	Geographical locations and landmarks.	<i>space, surface, street</i>
MEDIA	Various forms of communication and entertainment media.	<i>newspaper, track, publication</i>
MONEY	Monetary units, currencies, and financial values used in different contexts.	<i>dollar, 15 euros</i>
ORG	Organizations and institutions.	<i>Arsenal, Google</i>
PART	Individual components or sections of larger entities or objects.	<i>finger, chin, head</i>
PER	Individuals or historical figures.	<i>doctor, historian, professor</i>
PLANT	Types of trees, flowers, and other plants.	<i>grass, peach tree, Forsythia</i>
PROPERTY	Properties or attributes of objects, entities, or concepts.	<i>thickness, height, dimension</i>
RELATION	Relationships, connections, and associations between entities or concepts.	<i>apport, competition</i>
STRUCT	Physical structures or constructions.	<i>shelter, gravestone, refuge</i>
SUBSTANCE	Chemical substances.	<i>acid, bactericide, carbonyl</i>
SUPER	Mythological and religious entities.	<i>Apollo, Persephone</i>

Table 2.1. Label, description, and instance examples of each of our CNER categories.

	Concepts Semantic (Classes) Categories		CNER Categories
Food-101	102	4	FOOD, BIO, PLANT, SUBSTANCE
CIFAR-10	10	2	ANIMAL, ARTIFACT
Birdsnap	500	1	ANIMAL
SUN397	397	3	LOC, STRUCT, ORG
Stanford Cars	196	1	ARTIFACT
FGVC Aircraft	100	1	ARTIFACT
Pascal VOC 2007 Classification	20	4	ANIMAL, ARTIFACT, PER, PLANT
Describable Textures	47	1	PROPERTY
Oxford-IIIT Pets	37	1	ANIMAL
Caltech-101	102	4	ANIMAL, ARTIFACT, PLANT, PER
Oxford Flowers 102	102	1	PLANT
MNIST	10	1	MEASURE
Facial Emotion Recognition 2013	8	1	FEELING
STL-10	10	2	ANIMAL, ARTIFACT
EuroSAT	10	3	PLANT, STRUCT, LOC
GTSRB	43	3	ARTIFACT, MEASURE, LAW
PatchCamelyon	2	2	BIO, DISEASE
UCF101	101	3	PER, DISCIPLINE, CULTURE
Kinetics700	700	3	PER, DISCIPLINE, CULTURE
CLEVR Counts	8	2	MEASURE, ARTIFACT
Rendered SST2	2	2	PER, CULTURE
Hateful Memes	2	4	PER, CULTURE, ORG, LAW
ObjectNet	312	1	ARTIFACT
iNaturalist	8142	3	ANIMAL, PLANT, BIO
KITTI	4	5	ARTIFACT, PER, STRUCT, LOC, MEASURE
RESISC45	45	6	LOC, STRUCT, ARTIFACT, PLANT, ORG, CULTURE
Country211	211	6	LOC, STRUCT, CULTURE, PLANT, ANIMAL, PER
CIFAR-100	100	7	ANIMAL, PLANT, FOOD, ARTIFACT, STRUCT, LOC, PER
ImageNet-1K	1000	11	ANIMAL, ARTIFACT, BIO, STRUCT, FOOD, LOC, MEDIA, PER, PLANT, SUB, PROPERTY
Total (combined)	-	18	

Table 2.2. Comparison of concept and semantic category coverage for existing benchmarks. We report the number of unique concepts and semantic categories for each dataset, as well as the total number of distinct semantic categories covered across all the benchmarks (**Total (combined)**), including those in the CLIP Benchmark.

Chapter 3

Addressing the Gap in Multimodal Knowledge Coverage

3.1 Overview

This chapter addresses the first major objective of this dissertation, improving multimodal models’ ability to capture the breadth and diversity of real-world concepts. Building on the resources, tasks, and evaluation challenges introduced in Chapter 2, we focus here on a particularly consequential failure mode, namely limited semantic coverage in Vision-language Models (VLMs). While recent VLMs have achieved impressive performance on standard benchmarks, existing evaluation paradigms remain heavily concentrated on a narrow subset of semantic categories, primarily centered around concrete object recognition. As a result, current benchmarks provide only a partial view of a model’s ability to capture the diversity of real-world concepts and systematically obscure gaps in semantic coverage and generalization.

We argue that broad and structured concept coverage is essential for knowledge-intensive multimodal systems. Concepts constitute the semantic abstractions through which humans organize objects, attributes, events, and phenomena; robust concept understanding requires models to move beyond coarse category recognition toward fine-grained and domain-specific distinctions. This challenge becomes increasingly pronounced as the semantic space expands to include scientific, cultural, abstract, and long-tail concepts that are largely absent from widely used benchmarks.

To address these limitations, this chapter introduces approaches for automatically constructing large-scale resources and evaluation benchmarks that expand the semantic scope of multimodal learning and assessment. More specifically, the main contributions of this chapter are as follows:

- We introduce *Concept-pedia*, a large-scale multimodal resource grounded in Wikipedia and enriched with semantic annotations, enabling wide-coverage supervision across diverse concept types and domains.
- Building on this resource, we present *Concept-10k*, a manually validated benchmark that substantially expands the number and diversity of semantic categories used to evaluate visual concept recognition, providing a more representative view of real-world conceptual variability than standard benchmarks.
- Through extensive zero-shot and transfer evaluations, we demonstrate that state-of-the-art VLMs exhibit substantial performance degradation when confronted with the broader semantic diversity captured by Concept-10k, revealing systematic limitations that remain hidden under ImageNet-centric evaluation.
- We further show that Concept-pedia can be used to improve model generalization to previously unseen concept categories, yielding strong gains while preserving performance on established benchmarks.

We emphasize that the goal of this chapter is not to replace existing datasets and benchmarks, but to complement them by exposing dimensions of semantic coverage that are currently underrepresented. By establishing wide-coverage resources and evaluation protocols at the concept level, this chapter lays the foundation for the subsequent shift toward more entity-centric multimodal knowledge in the next chapter.

3.2 Vision-Language Concept Recognition

We previously highlighted a central limitation of current multimodal evaluation: despite strong performance on standard benchmarks, existing VLMs are assessed on a narrow and saturated subset of the conceptual space. Here, we elaborate on why this limitation arises,

why it matters for knowledge-intensive multimodal systems, and how it motivates the need for wide-coverage concept resources.

Concepts as the Basis of Multimodal Knowledge. Concepts constitute the semantic abstractions through which both humans and machines interpret the world. High-level categories such as *object*, *animal*, or *vehicle* capture broad similarities, but effective multimodal understanding also requires knowledge about finer-grained and domain-specific distinctions. For example, while both a *Japanese macaque* and a *crab-eating macaque* fall under the broader category of *monkey*, distinguishing between them depends on semantic knowledge that goes beyond coarse category membership, including species-specific attributes and differences. These distinctions reflect different levels of semantic granularity that are essential for knowledge-intensive reasoning over both images and text.

Concept Representation in VLMs. Recent VLMs have achieved remarkable success by learning joint representations of images and text from large-scale image-text corpora. Models such as CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) employ contrastive objectives over web-scale data, enabling strong zero-shot performance on widely used visual concept recognition benchmarks. Several works have shown that both the scale and the curation of training data play a decisive role in these capabilities, as demonstrated by analyses of large pre-training corpora such as LAION (Schuhmann et al., 2022) and DataComp (Gadre et al., 2023). These findings suggest that modern models implicitly acquire knowledge about a wide range of concepts during pre-training.

ImageNet-Centric Evaluation. Despite these advances, the evaluation of visual concept recognition remains tightly centered on a small number of widely used benchmarks. Datasets such as ImageNet-1k (Deng et al., 2009), ObjectNet (Barbu et al., 2019), and iNaturalist (Horn et al., 2018) have played a foundational role in tracking progress. Among them, ImageNet-1k (see Section 2.1.3) has emerged as the de facto general-purpose benchmark due to its relative diversity across objects, animals, food, and artifacts. However, this apparent breadth masks a structural limitation: ImageNet-1k covers only 1,000 concepts, while the current complete ImageNet taxonomy contains 21,841 concepts, representing only

a fraction of real-world concepts (just 26% of all WordNet nominal synsets).¹.

This reliance on ImageNet introduces two well-documented issues. First, performance improvements often reflect increasing specialization to its specific image and label distributions rather than robust semantic generalization (Recht et al., 2019; Beyer et al., 2020). Second, existing resources and benchmarks are heavily skewed toward concrete object categories, systematically underrepresenting abstract concepts, properties, scientific notions, cultural artifacts, and other concepts that frequently appear in real-world multimodal content. Variants such as ImageNet-v2, ImageNet-R, ImageNet-A, and ImageNet-ReaL address distribution shift or annotation quality, but remain constrained to the same underlying concept set (Recht et al., 2019; Hendrycks et al., 2021a,b; Beyer et al., 2020). Similarly, ObjectNet and iNaturalist expand along specific dimensions, yet still cover a limited diversity of the concepts they consider. These limitations motivate the need for multimodal resources and evaluation benchmarks that substantially expand semantic coverage, moving closer to real-world concept representation.

Building on Wikipedia’s role as a core grounding substrate (Chapter 2), we ground concept annotation in Wikipedia to leverage its semantic structure and curated links. While, as previously mentioned, Wikipedia hyperlinks have long been used to construct large resources for text-only NLP tasks (Nothman et al., 2013; Tsai et al., 2016; Pan et al., 2017; Wu et al., 2020; Conia et al., 2024), we use them here to build a wide-coverage multimodal resource for visual concept recognition. This provides the foundation for the methodology introduced in the next section, where we describe the creation of *Concept-pedia* and its use in evaluating VLMs across a broader and more representative semantic space.

3.3 Concept-pedia

We now detail how images from Wikipedia are associated with their respective concepts. We use hyperlinks in image captions as the basis for mapping images to concepts with the help of BabelNet² (Navigli and Ponzetto, 2010; Navigli et al., 2021), a multilingual knowledge graph. However, since Wikipedia is manually edited, many captions lack direct hyperlinks. To address this limitation, we automatically annotate the captions with the missing links.

¹<https://www.image-net.org/>

²<https://babelnet.org>

Next, we generalize entities to the concepts they are instances of (e.g. *Colosseum* is an instance of amphitheater), and ensure that each concept in our resource is associated with an adequate number of images by exploiting BabelNet’s taxonomical information. Striking this balance is crucial to maintaining the dataset’s quality and representativeness.

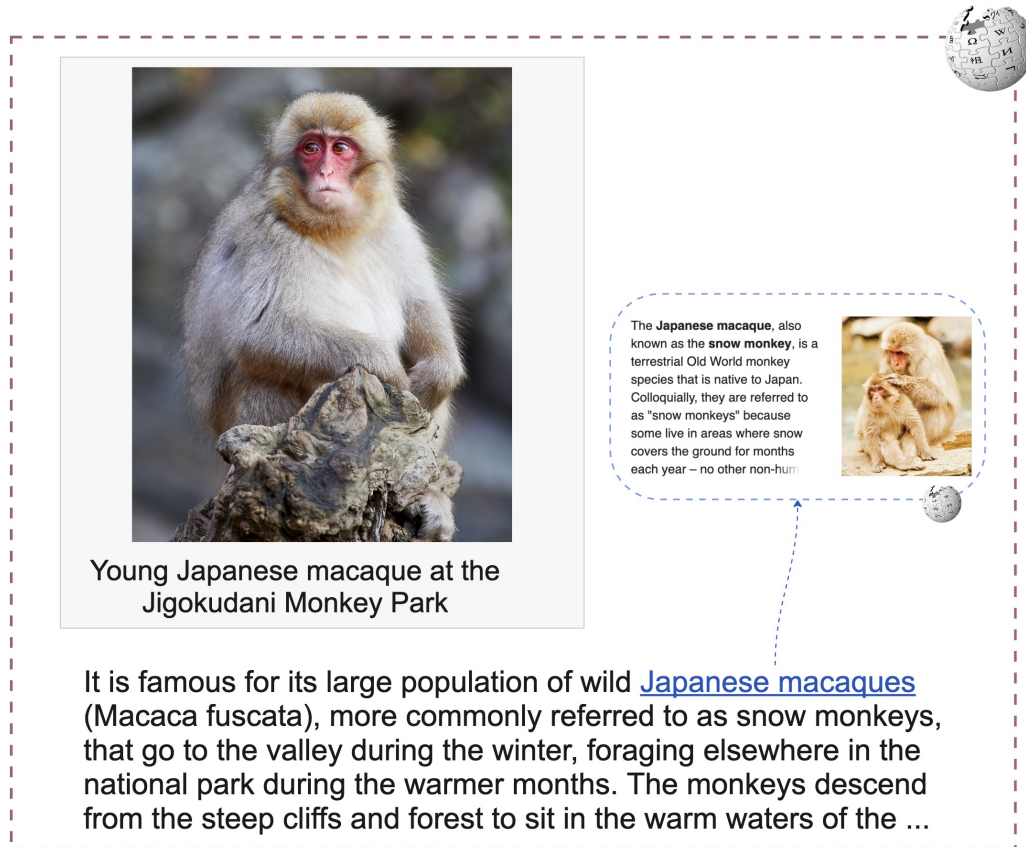


Figure 3.1. An example of a Wikipedia page containing an image-caption pair in which the caption lacks a hyperlink to the mentioned concept, while the page body contains the corresponding link.

3.3.1 Link propagation

Wikipedia pages provide manually annotated hyperlinks, including those found in image captions. However, many mentions of entities or concepts within captions lack corresponding hyperlinks, resulting in missed annotations, as illustrated in Figure 3.1. To overcome this issue, we employ a link propagation algorithm inspired by Tedeschi et al. (2021), which matches unannotated text chunks in a caption with links present in the Wikipedia page containing the caption.

The algorithm operates on the assumption that captions frequently reference mentions already linked within the body of the corresponding Wikipedia page. Building on this assumption, the propagation algorithm attempts to match unlinked text chunks in captions with manually annotated links in the page. This matching process relies on a sequence of cascaded heuristics to expand the annotations. First, **exact string matching** identifies exact surface form overlaps between text chunks and linked mentions, ensuring high precision when captions use identical terminology. Then, **lemma matching** extends this by normalizing text to its lemmatized form, which we perform using SpaCy³, enabling matches that account for inflectional variations such as pluralization or verb conjugation (e.g., *cathedrals* being matched with *cathedral*).

To further link more unannotated text chunks, we propose two heuristics that utilize the inter-resource mapping provided by BabelNet to match text chunks with synonyms or alternative lexicalizations. The first heuristic, referred to as **synonym matching**, obtains for each existing link in a Wikipedia page the corresponding BabelNet synset, and identifies unannotated text chunks that match WordNet synonyms or Wikidata aliases contained within that synset. For instance, *natatorium* can be linked to the *swimming pool* page since both terms are synonyms in the corresponding WordNet synset, and *gas* linked to *gasoline* due to the former being a Wikidata alias. This technique is particularly useful for handling variations in naming conventions, such as acronyms or synonyms. The second heuristic, i.e. **redirection matching**, performs an analogous matching using Wikipedia page redirections, i.e. pages that redirect to other pages. This allows us to capture alternative spellings, common misspellings, and other variations, e.g. *Red Panda* and *Lesser Panda*.

While the link propagation algorithm significantly enhances coverage, the precision of its heuristics varies. Hyperlinks manually added to captions naturally exhibit the highest precision, as they were explicitly added by Wikipedia editors. In contrast, heuristics like synonym or redirection matching introduce a degree of uncertainty, e.g. potentially introducing concept shifts (e.g. *Driver (person)* redirecting to *Driving*). To maximize coverage while maintaining transparency, Concept-pedia includes annotations derived from all heuristics and records the link propagation information for each link. This enables downstream applications to tailor their use of the annotations based on their needs – prioritizing precision

³<https://spacy.io/>

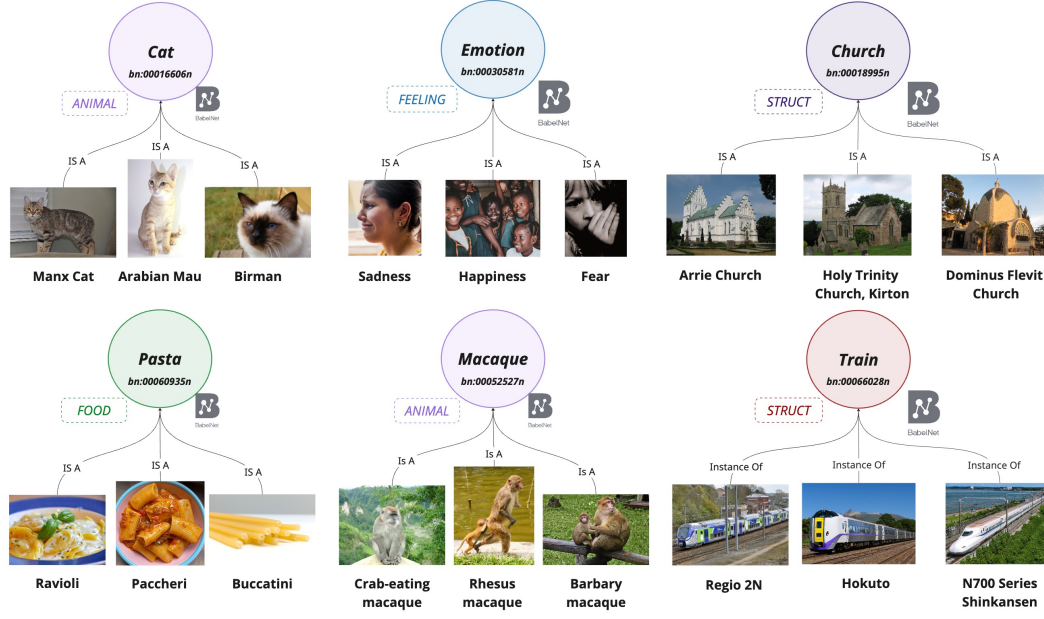


Figure 3.2. Examples for taxonomical concept population in Concept-pedia across different CNER categories.

(e.g., by using only existing or exact matches) or maximizing recall (e.g., by incorporating all annotations). We discuss this in more detail in Section 3.5.

This link propagation approach to creating multimodal datasets achieves two key objectives. First, creating manually annotated datasets is notoriously expensive; this method repurposes high-quality manual annotations, thereby significantly reducing costs. Second, as highlighted by Beyer et al. (2020), images inherently contain multiple concepts, which our link propagation algorithm helps associate. By leveraging this approach, we produce what is, to the best of our knowledge, the first semantically-annotated multimodal training resource of this scale with multi-label concept annotations.

3.3.2 Taxonomical Concept Propagation

After the link propagation step, 74.5% of the 4.6 million English Wikipedia captions (Nov, 2023 dump) are annotated with at least one link. We leverage the correspondence established in BabelNet between its synsets and Wikipedia links to map these annotations. This enables us to exploit BabelNet’s synset classification into concepts (e.g. *City*) and named entities (e.g. *Paris*), revealing that approximately 85% of the synsets involved are classified as named entities.

To maximize the size of our dataset and maintain a focus on concepts rather than specific entities, we transform each entity annotation into its corresponding generalized concept using BabelNet’s taxonomic information, which is derived from the integration of WordNet and Wikidata hypernymy and instance-of relations. For example, the entity *Arrie Church* is generalized to the concept *Church* (see Figure 3.2). Additionally, we observe that many concepts are associated with only a few images (see Table 3.1, last two columns). To ensure that each concept is adequately represented, we iteratively refine the annotations for underrepresented synsets, i.e. those synsets with fewer images than a predetermined threshold. To do so, we map each annotation to its direct parent concept (e.g., mapping *Hokuto* to *Train*). This process is repeated until every concept meets the minimum image requirement, ensuring a balanced distribution across the training, development, and test splits. An additional benefit of grounding our dataset in BabelNet is that its taxonomy preserves a hierarchical structure, similarly to ImageNet in this respect, while covering a broader range of synsets beyond those available in WordNet.

Model-based Filtering It is important to note that we opted against using model-based selection methods (e.g., CLIPScore (Hessel et al., 2021)). There are two main reasons behind this decision. First, such methods inherently introduce bias (Gadre et al., 2023). Second, we qualitatively observe that current models struggle to accurately assess the similarity between the visual and textual representations of new concepts, often resulting in the erroneous exclusion of valid samples.

Propagation	Capt.	Links	Concept	
			Unique	Freq.
w/o	1.1M	0.6M	114K	46K
+ Link	3.4M	1.2M	165K	48K
+ Taxonomical	3.4M	1.2M	101K	101K

Table 3.1. Statistics of Concept-pedia at each stage of the propagation algorithm, showing the number of captions with at least one link (Capt.), unique Wikipedia link annotations (Links), unique concepts (Unique), and frequent concepts that appear more than 5 times (Freq.) at each step.

3.3.3 Dataset Statistics

Table 3.1 presents Concept-pedia statistics at each step, including the number of captions, links, and concept annotations. Additionally, we report the number of concepts that occur at least five times, referred to as frequent concepts.

We observe that, while the Link Propagation step (Section 3.3.1) doubles the number of links (row 2 of Table 3.1) and adds approximately 50,000 new concepts, it has little effect on frequent concepts, i.e., concepts appearing more than 5 times. In contrast, the Taxonomical Propagation step (Section 3.3.2) flattens the set of unique concepts to 101,354 and ensures that each of them is associated with at least 5 images in the dataset (row 3 of Table 3.1). Finally, we compare ImageNet-21k, another large-scale resource, to Concept-pedia in terms of category coverage and distribution, as illustrated in Figure 3.3.

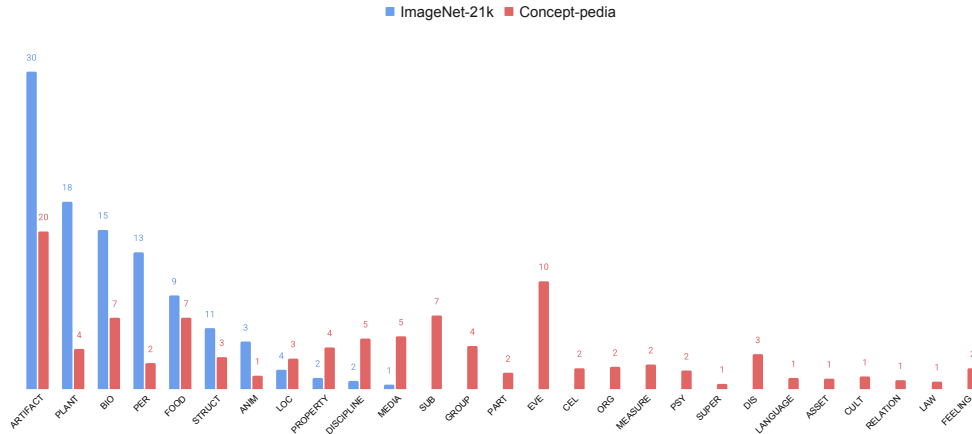


Figure 3.3. Comparison of CNER category population between Concept-pedia and ImageNet-21k.

3.4 Concept-10k

Having created a large silver dataset with Concept-pedia, we can proceed to develop a gold-standard benchmark. Based on the results of our manual validation, presented in Section 3.5, we observed that larger captions or those containing multiple concepts are more prone to propagating concepts that are not visually represented in the images. To address this, starting with concepts that have at least 10 images, we apply three heuristics to Concept-pedia to

select the most representative image-concept pairs: 1) we filter out all image-caption pairs with multiple concepts, leaving their inclusion to future work; 2) for each concept, we rank images by caption length and retain the k shortest-caption images (we set $k = 5$); 3) we remove any concepts that are parents of other concepts according to the BabelNet taxonomy, e.g., if we have *Car* and *Race car*, we remove *Car* to avoid false negatives at test time.

The final step in creating our Concept-10k gold-standard benchmark involves manually validating each image-concept pair. In this task, two expert linguists⁴ with extensive experience in annotating lexical-semantic tasks were presented with image-concept pairs and instructed to perform a binary classification (True or False) based on whether the image was labeled with the correct concept. Additionally, the annotators were tasked with filtering out any Not Safe For Work (NSFW) or corrupt images. The inter-annotator agreement for this task was measured, yielding a Cohen’s kappa coefficient of $\kappa = 89.3$, indicating almost perfect agreement. Further details regarding this annotation process can be found in Section 3.5.4.

These heuristics are intended to produce a controlled gold-standard benchmark for this first iteration of Concept-10k, rather than to assume that real-world visual scenes contain only a single concept. In particular, filtering image-caption pairs with multiple concepts reduces ambiguity about which visual evidence supports the target label, making it easier to attribute model successes and failures to the intended concept. This design also preserves comparability with widely used concept-recognition benchmarks such as ImageNet-1K, where each image is associated with a single target concept, and thus provides a controlled reference point for assessing the quality of our largely automatic data creation pipeline. At the same time, we recognize that multi-concept scenes are central to realistic visual understanding; extending Concept-pedia to explicitly model such cases through multi-label annotations, region-level grounding, or structured concept representations is an important direction for future work.

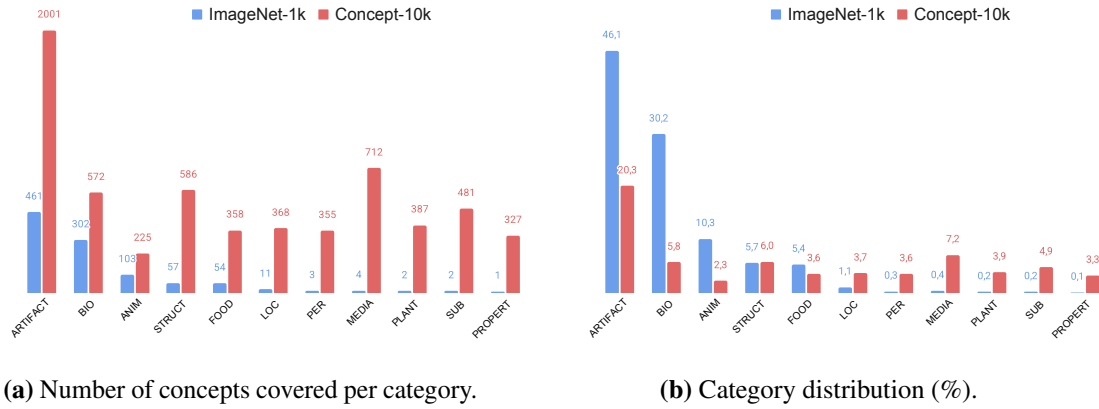
3.4.1 Benchmark Statistics and Analysis

The above procedure results in Concept-10k, a manually validated benchmark of 49,185 image-concept pairs, spanning 9,837 unique concepts, each represented by five images. To

⁴Annotators were compensated according to the standard salary for their geographical location.

	Concepts	Categories
ImageNet-1k	1,000	11
ObjectNet	312	1
iNaturalist	8,142	3
Concept-10k	9,837	28

Table 3.2. Number of unique concepts and semantic categories covered in each dataset.



(a) Number of concepts covered per category.

(b) Category distribution (%).

Figure 3.4. Concept and category distributions of Concept-10k and ImageNet-1k.

analyze the domain coverage of our final benchmark, we utilize the Concept and Named Entity Recognition (CNER) category set introduced by Martinelli et al. (2024), which maps BabelNet synsets to a predefined set of 29 semantic categories shared across concepts and entities. This mapping enables a systematic comparison between our resource and ImageNet-1k, enabling an in-depth analysis of the category coverage of the two resources. We provide Concept-pedia instances of images and concepts from CNER categories in Table 3.3 and Figure 3.5.

ImageNet-1k spans only 11 out of 29 CNER categories, as shown in Table 3.2, with just 6 containing more than four distinct concepts. In contrast, Concept-10k covers 28 categories, excluding only *DateTime*, and includes a significantly higher number of concepts for the categories it shares with ImageNet-1k, as shown in Figure 3.4a. This demonstrates that Concept-10k enables a broader evaluation, encompassing not only concrete concepts but also abstract ones, such as *Feelings* or *Events*, which are often underrepresented in existing benchmarks. A comparison with other benchmarks, including those from Radford et al. (2021), is provided in Section 3.5.3, showing that Concept-10k introduces more

categories (28) than all other benchmarks combined (18). Furthermore, as shown in Figure 3.4b, not only does Concept-10k cover more concepts per category than ImageNet-1k but it also exhibits a less biased distribution toward categories such as *Artifacts*, as seen in ImageNet-1k.

Category	Examples
ANIMAL	<i>Lion, American bison, Orca, Leopard, Reindeer</i>
ARTIFACT	<i>Racing car, Water tower, Wind turbine, Arcade video game</i>
ASSET	<i>Stock, Debit card, Mortgage loan, Equity, Health insurance</i>
BIOLOGY	<i>Proton, Protein dimer, Collagen</i>
CELESTIAL	<i>Pulsar, Star, Ring galaxy</i>
CULTURE	<i>Humanism, Sufism, Islam, Pacifism</i>
DISEASE	<i>Covid-19, Tuberculosis, Malaria, Smallpox, Ebola</i>
DISCIPLINE	<i>College football, Professional boxing, Contemporary art</i>
EVENT	<i>Freestyle swimming, Climate change, Deforestation</i>
FEELING	<i>Love, Nirvana, Hatred, Pride, Sympathy</i>
FOOD	<i>Chocolate bar, Honey, Mussels, Sushi, Ramen</i>
GROUP	<i>Brigade, Basketball team, Opera company</i>
LANGUAGE	<i>Ambigram, Synonym, Vowel</i>
LAW	<i>International law, Regulation, Independence</i>
LOC	<i>Vineyards, Playground, Bus stop</i>
MEASURE	<i>Kilometre, Latitude, Mile</i>
MEDIA	<i>Music video, Self-portrait, Visual novel</i>
MONEY	<i>Euro, Dirham, Dollar,</i>
ORG	<i>Music industry, Democratic Party, Commercial enterprise</i>
PART	<i>Liver, Stomach, Wrist, Retina</i>
PER	<i>Gymnast, Jockey, Ski jumper, Cyclist</i>
PLANT	<i>Bamboo, Sugarcane, Bonsai, Wheat</i>
PROPERTY	<i>Linear, Dimension, Shape</i>
PSYCH	<i>Sweetness, Cognition, Attention, Necessity</i>
RELATION	<i>approximation, Relatedness, Distance (graph theory), Function</i>
STRUCT	<i>Supermarket, Marina, Hospital</i>
SUBSTANCE	<i>Aluminium, Silver, Sodium</i>
SUPER	<i>Unicorn, Zombie, Goblin</i>

Table 3.3. Frequent concepts from Concept-pedia categorized by CNER groups.

3.5 Dataset Evaluation

To evaluate the quality of Concept-pedia and Concept-10k, we conduct a series of manual annotation experiments.

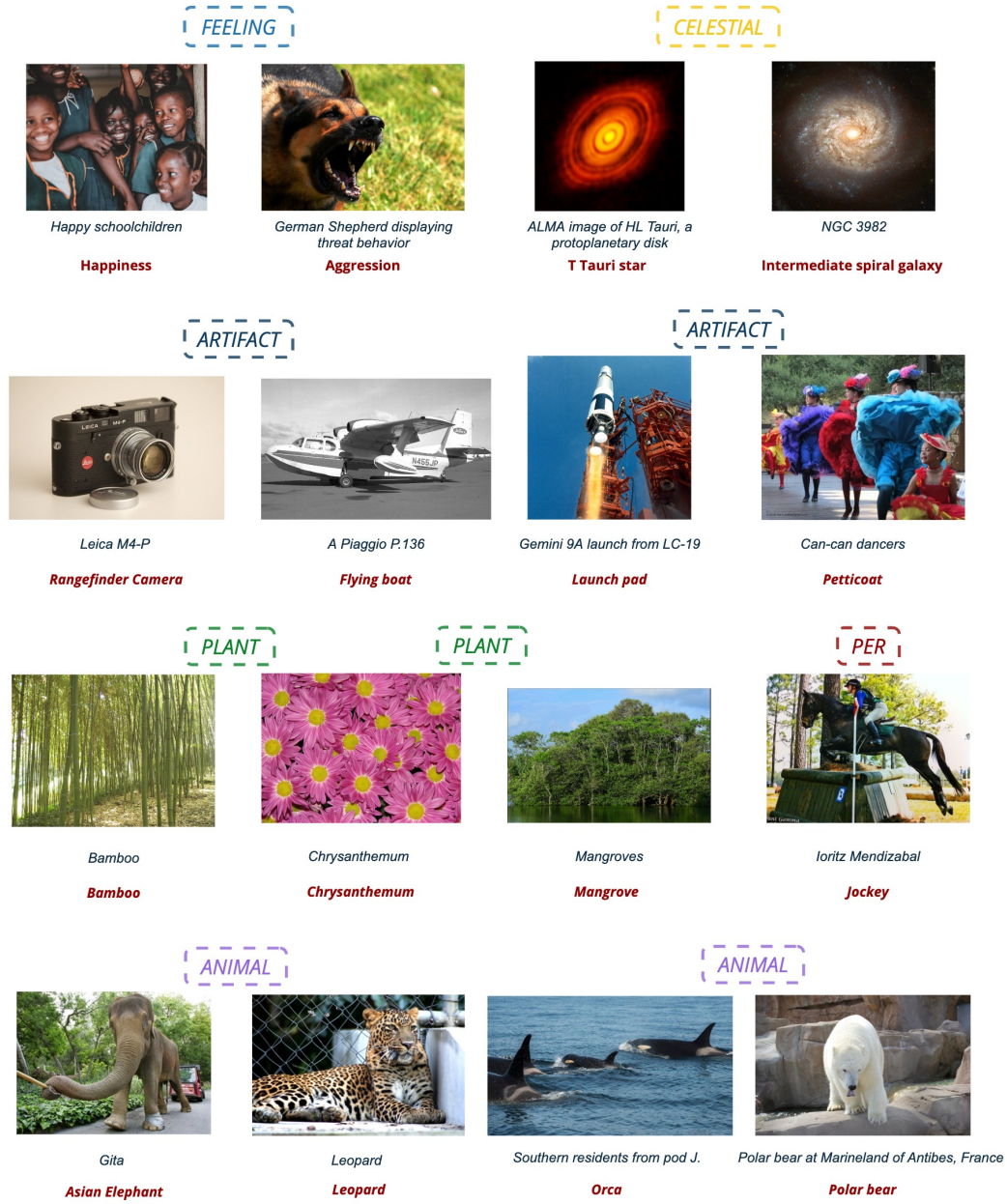


Figure 3.5. Examples from Concept-pedia. We provide the caption (in blue) and the concept (in red) for each image.

3.5.1 Annotation Quality Assessment

We first evaluate the effectiveness of the Link Propagation step in our Concept-pedia creation methodology (Section 3.3.1). To do so, we randomly sample 200 image-caption-link instances for each annotation type (i.e., source links and the heuristics described) and manually evaluate them. As shown in Table 3.4, precision is generally high, with all

heuristics exceeding 90%, except for redirection matching, which, as expected, yields slightly lower precision. For the Taxonomical Concept Propagation step, we evaluate 300 randomly sampled links generalized to concepts, and observe a precision of 96%.

Propagation Type	P
Hyperlink in caption	100
Exact string matching	100
Lemma matching	91
Synonym matching	96
Redirection matching	84
Taxonomical propagation	96

Table 3.4. Precision (%) scores for the link and taxonomical concept propagation heuristics.

3.5.2 Comparison with ImageNet-1k

To further assess the quality of our silver dataset, we focus on two key aspects: the accuracy of the images in representing the given concepts and the precision of our automatic annotations compared to existing benchmarks, such as ImageNet. To this end, we conduct an experiment using an equal number of images from ImageNet and a restricted version of Concept-pedia that contains only ImageNet-1k concepts, referred to as Concept-1k. We evaluate 2,000 examples by randomly sampling two images per concept from both Concept-1k and ImageNet-1k, ensuring equal representation across concepts. This setup enables a fair comparison of quality between our dataset and a well-established benchmark.

Two expert linguists, following the guidelines detailed in Section 3.5.4, are provided with 10 possible candidate labels, i.e., Wikipedia titles, for each image, retrieved from different vision-text models⁵, and instructed to select all the prominent concepts depicted in the image. To ensure unbiased evaluation, instances are shuffled across resources and

⁵We use SigLIP-base, SigLIP-large, and SigLIP-(SO)-400M, CLIP-base and CLIP-large and select the 10 most predicted labels.

	ImageNet-1k	Concept-1k
High	88.4%	95.7%
Low	83.7%	76.8%

Table 3.5. Label accuracy of manually annotated subsets of ImageNet-1k and Concept-1k.

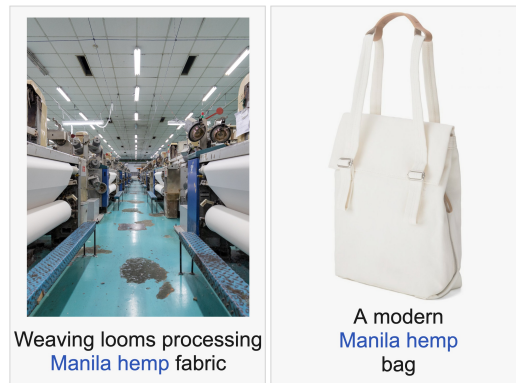


Figure 3.6. Examples of correct annotations excluded from Concept-10k. This is due to the concept not being visually represented in the image.

are shown without any source information. We accept *None* as a possible label to identify instances where the concept is correctly tagged given the context in the caption but not visually represented in the image, as shown in Figure 3.6. Given the fine-grained nature of ImageNet-1K classes (e.g., specific dog breeds), annotators are also asked to rate their confidence in their labels as High or Low. This process results in 84% of items in Concept-1K and 89% in ImageNet-1K being tagged with high confidence by both annotators.

For these items, as shown in Table 3.5, Concept-1k labels are confirmed as correct 95.7% of the time, compared to 88.4% for ImageNet-1k, validating the quality of our automatic labels and highlighting our dataset’s competitiveness with a well-established benchmark. The inter-annotator agreement, measured on a shared subset of 200 instances, yields a Cohen’s kappa coefficient of $\kappa = 88.1$, indicating almost perfect agreement.

3.5.3 Benchmark Coverage Comparison

Comparison with ImageNet-CoG As previously mentioned, ImageNet-CoG was created to provide concepts that are not included in ImageNet. In Figures 3.7 and 3.8, we provide details on the distribution of the remaining categories in Concept-10k, along with a comparison against ImageNet-CoG in terms of CNER category coverage and population. Even though ImageNet-CoG includes more concept categories than ImageNet-1k, we can clearly see that it suffers from a similarly skewed category distribution toward Artifacts.

Benchmark Category Coverage In Table 3.6, we report the concept and semantic category coverage of commonly used benchmarks, including ImageNet-1k, iNaturalist, Object-

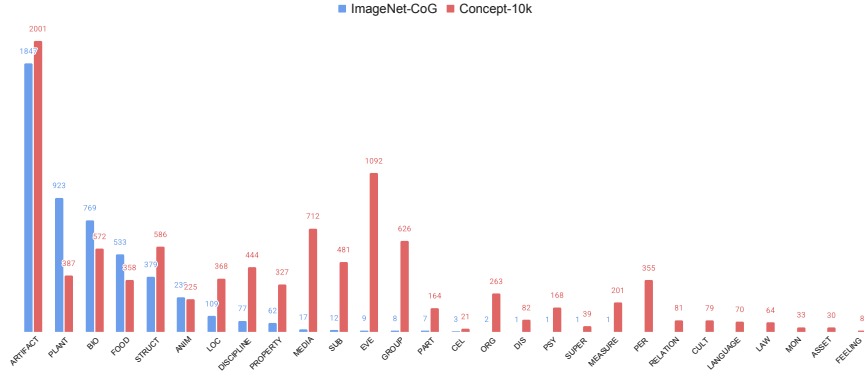


Figure 3.7. Number of concepts per CNER category in Concept-10k and ImageNet-CoG.

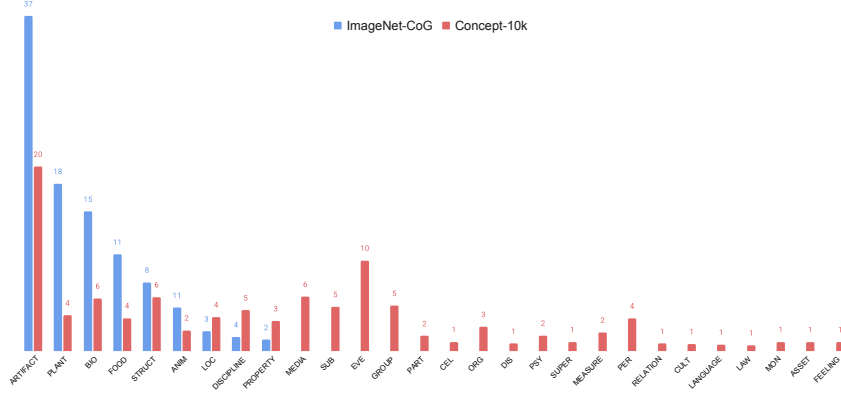


Figure 3.8. CNER category distribution (%) of Concept-10k and ImageNet-CoG.

Net, and the 27 datasets used in (Radford et al., 2021), and compare them with Concept-10k. We show that Concept-10k not only includes significantly more concepts than the other benchmarks, but also covers a broader range of semantic categories. Notably, even when combining all other benchmarks, the total number of unique semantic categories amounts to 18, which Concept-10k surpasses with 28 categories.

3.5.4 Annotation Guidelines

In this section, we outline the annotation guidelines employed in our study. We assign each annotation task to two expert linguists. Annotators were paid according to the standard salaries for their geographical location. Annotators were tasked with performing three distinct annotation tasks to evaluate the automatic annotations and the associated images. Due to the different requirements of these tasks, we developed a custom application for each

	Concepts (Classes)	Semantic Categories
Food-101	102	4
CIFAR-10	10	2
Birdsnap	500	1
SUN397	397	3
Stanford Cars	196	1
FGVC Aircraft	100	1
Pascal VOC 2007 Classification	20	4
Describable Textures	47	1
Oxford-IIIT Pets	37	1
Caltech-101	102	4
Oxford Flowers 102	102	1
MNIST	10	1
Facial Emotion Recognition 2013	8	1
STL-10	10	2
EuroSAT	10	3
GTSRB	43	3
PatchCamelyon	2	2
UCF101	101	3
Kinetics700	700	3
CLEVR Counts	8	2
Rendered SST2	2	2
Hateful Memes	2	4
ObjectNet	312	1
iNaturalist	8142	3
KITTI	4	5
RESISC45	45	6
Country211	211	6
CIFAR-100	100	7
ImageNet-1k	1000	11
Total (combined)	-	18
Concept-10k	9837	28

Table 3.6. Comparison of concept and semantic category coverage between Concept-10K and existing benchmarks. We report the number of unique concepts and semantic categories for each dataset, as well as the total number of distinct semantic categories covered across all the benchmarks (**Total (combined)**), including those in the CLIP Benchmark. Concept-10k covers all CNER classes except *DateTime*.

in order to facilitate the process.

Annotators received detailed, task-specific guidelines, which we describe below. As a general guideline for all tasks, annotators were instructed to discard corrupt or unclear images, as well as NSFW images and concepts.

Furthermore, for every concept utilized across tasks, we provided the annotators with the definition retrieved from different resources through BabelNet, e.g. Wikidata, WordNet. In case of persistent ambiguity occurring with a given concept, if the human annotator could not determine the correct annotation, with a high degree of confidence, they were required to flag the given sample.

Task 1: Silver Labels Evaluation

Task Description In this task, annotators are presented with image-caption pairs where hyperlinks are visible and clickable, as they appear in Wikipedia. We supplement the existing hyperlinks in the captions with additional links provided by our link propagation approach, as described in 3.3.1. Annotators are required to click on the provided hyperlink in the caption and verify whether it corresponds to it and is associated correctly. Some examples are provided in Figure 3.9.

Task 2: Concept-1k vs. ImageNet-1k

Task Description

1. Check what the main concept *could* be in the image. It needs to have visually clear features.
2. The image can be inside or outside of a building as long as it is visually clear. If an image clearly depicts a specific location or structure, such as the interior of a church with recognizable features like stained glass windows and pews, it should be tagged accordingly.
3. If the concept does not cover the *majority* of the image, then it is not considered the main concept.
4. The image needs to depict most of the concept. For example, an image focused only on an “altar”, which is peculiar to a church but not representing a church, can not be tagged as a “church”. On the contrary, an image with the inside of a church should be tagged as church because it is the biggest inner part of the church.



Figure 3.9. Examples of link propagation annotations in Concept-pedia.

5. When an object is the image's primary focus, and a person is present but not the main subject, prioritize tagging the object. For example, if an image shows a person kayaking on a lake, the appropriate tag would be "kayak" rather than "person", since the kayak is the specific object of interest.
6. In the case of two competing concepts, tag the image with both concepts.
7. The maximum number of tags should be two. If there are more than two, tag the image as **uncertain** and provide an explanation in the notes tab.
8. If none of the concepts provided correctly describe the image, leave the sample untagged, and mention it in the notes with the probable main concept in your opinion.
9. Kindly use the notes tab extensively to signal any issues regarding the process or the data.



“Church”

“Altar”



Examples

- **Example 1: Clear Single Concept**

- **Image Description:** A close-up photograph of a sunflower in a field, filling the entire frame with its petals and center clearly visible.
- **Tag:** "sunflower"
- **Motivation:** The sunflower is the dominant and sole subject of the image, with no other competing elements.

- **Example 2: Specific Object with a Person Present**

- **Image Description:** An image of a person kayaking on a serene lake, with the kayak being prominently visible and the person facing away from the camera.
- **Tag:** "kayak"

- **Motivation:** The kayak is the main focus, and while a person is present, they are not the primary subject. The emphasis is on the activity and the kayak itself.

- **Example 3: Two Competing Concepts**

- **Image Description:** A dog and a cat sitting side by side on a couch, both looking directly at the camera and occupying equal space in the frame.
- **Tags:** "dog," "cat"
- **Motivation:** Both animals are equally prominent and central to the image, making it appropriate to tag both.

Task 3: Concept-10k Validation

We followed the instructions outlined in (Krizhevsky and Hinton, 2009), with one key difference: we provided definitions for each concept, extracted from sources such as Wikidata, Wikipedia, and WordNet, to eliminate any potential ambiguity. Annotators were tasked with selecting the images to be included in the final benchmark. This step was essential because, in some cases, the label propagated by our algorithm was correct in the context of the caption but the concept was not visually represented in the image.

In Figure 3.10, we present two such examples. The propagated label is shown in red, while the caption is shown in blue. Although the label is accurate relative to the caption, it is not suitable as a Visual Concept Recognition label.

3.6 Experiments and Results

3.6.1 Experimental setup

We assess whether current models can generalize to new concepts beyond those included in ImageNet. To achieve this, we evaluate a diverse set of open-source vision-text transformer-based models, using them as baselines in our experiments and analyzing their accuracy on standard variations of ImageNet-1k (i.e., ImageNet-1k Val, ImageNet-v2, and ImageNet-Real) as well as on Concept-10k. Our evaluation includes the SigLIP family of models

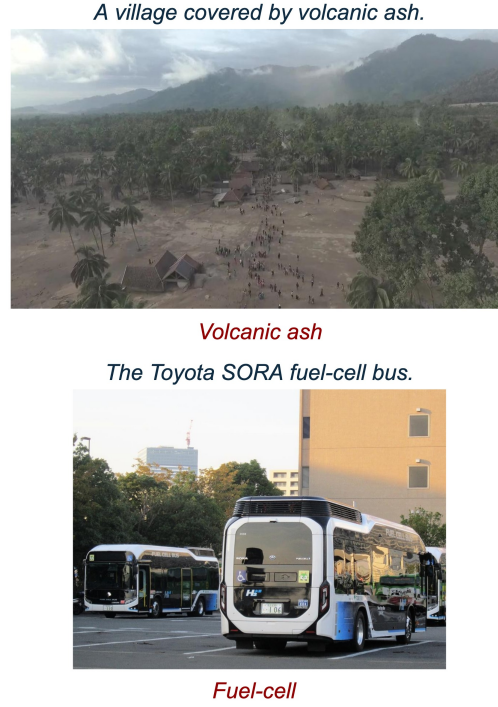


Figure 3.10. Filtered negative examples from Concept-10k.

(Zhai et al., 2023), featuring three configurations: SigLIP-base, SigLIP-large, and SigLIP with a Shape-Optimized (SO) 400M parameter Vision Transformer (ViT) architecture. Additionally, we evaluate two models of the CLIP family (Radford et al., 2021), namely CLIP-base and CLIP-large.

To test the models’ ability to recognize concepts, we embed candidate concepts and input images into the same latent space and predict the class with the highest cosine similarity to the image embedding, and thus measure top-1 accuracy. This unified evaluation framework standardizes predictions across models, ensuring a fair and consistent comparison. Although SigLIP and CLIP are trained with different optimization objectives, as SigLIP employs a sigmoid-based loss function and CLIP utilizes a softmax loss, we apply this methodology to evaluate both models consistently. We carry out our evaluation in two scenarios, namely, zero-shot and fine-tuned, presented hereafter.

MLLM and VLM Evaluation We also evaluated the visual concept recognition capabilities of Multimodal Large Language Models (MLLMs) such as LLava-1.5 (Liu et al., 2024) and InstructBLIP (Dai et al., 2023b), both of which use a 7-billion-parameter LLM, as well

Model	ImageNet-1k	Concept-10k
CLIP-base	68.3	22.0
SigLIP-base	76.2	31.1
CLIP-large	75.5	26.7
SigLIP-large	80.5	36.1
SigLIP-so400m	83.2	40.3
FLAVA	36.2	4.8
LLava-1.5	22.8	2.8
InstructBLIP	14.6	2.6

Table 3.7. Comparison of zero-shot model performance (top-1 accuracy) on ImageNet-1k and Concept-10k.

	ImageNet-1k	Concept-1k
SigLIP-base	86.3	79.1
SigLIP-large	87.6	83.0

Table 3.8. Model prediction accuracy (%) on manually annotated subsets of ImageNet-1k and Concept-1k, evaluated against high-confidence manual annotation labels.

as other encoder-only models like FLAVA (Singh et al., 2022). Given that Concept-10k comprises approximately 10,000 classes and these models support a maximum context length of 4K tokens, it is infeasible to include all class names directly in the prompt. To address this, inspired by Cao et al. (2021), we adopt a constrained decoding approach that restricts the model’s output space to the set of candidate classes. However, under this evaluation setting, these models exhibit substantially poor performance, particularly on Concept-10k. For this reason, we focus the remainder of our experiments on CLIP and SigLIP, which indicate stronger and more stable performance for large-scale visual concept recognition.

3.6.2 Zero-shot Visual Concept Recognition

In this section, we present the zero-shot performance of the models evaluated on both ImageNet-1k variants and our novel benchmark, Concept-10k. The latter aims to provide insight into the ability of pre-trained models to scale on categories not included in ImageNet-1k. As shown in Table 3.9, models perform relatively well on ImageNet-1k concepts, but struggle to accurately recognize the wide range of concepts present in the Concept-10k test set across the board. Even SigLIP-SO400m, the best-performing model on all variations of ImageNet-1k, fails to exceed 40.3% accuracy on Concept-10k.

Model	ViT size	Params	Fine-Tuning	ImageNet-1k			Concept-10k
				Validation	v2	ReaL	
CLIP	B	150M	$\times$	68.3	61.9	-	22.0
SigLIP	B	203M	$\times$	76.2	69.6	82.8	31.1
SigLIP	B	203M	$\checkmark$	75.9	69.1	81.5	36.1
CLIP	L	428M	$\times$	75.5	69.0	-	26.7
SigLIP	L	652M	$\times$	80.5	74.2	85.9	36.1
SigLIP	L	652M	$\checkmark$	79.3	73.5	85.3	41.9
SigLIP	SO (400M)	878M	$\times$	83.2	77.2	87.5	40.3
SigLIP	SO (400M)	878M	$\checkmark$	83.0	76.7	87.1	45.0

Table 3.9. We report the zero-shot and fine-tuned performance of different models. Performance is measured using top-1 accuracy (acc@1) for all datasets, except for ImageNet ReaL, where we report ReaL accuracy instead. We also report the performance of SigLIP fine-tuned on Concept-pedia on ImageNet test sets as well as on Concept-10k.

To determine whether the introduced categories are responsible for the observed drop in performance, we evaluate model accuracy on Concept-1k, which includes only the concepts shared with ImageNet-1k Validation, as discussed in Section 3.5.2. As shown in Table 3.8, performance on Concept-1k decreases compared to ImageNet-1k. A similar performance drop is observed with ImageNet-v2 (cf. Table 3.9), which, like Concept-1k, introduces new images for ImageNet-1k concepts. These results suggest that the performance drop is primarily driven by the increased diversity of concepts rather than by the quality of the images or annotations.

3.6.3 Concept-pedia Fine-tuning

To determine the quality and usefulness of Concept-pedia in improving models’ ability to detect new concepts, we first filter out all Concept-10k images. We then fine-tune SigLIP on the resulting dataset using a contrastive learning approach. SigLIP is selected in our experiments for two main reasons: first, it is the best-performing model for its size, as shown in Table 3.9, and second, it is less sensitive to batch size compared to CLIP, making it a more practical choice given our academic budget constraints. As shown in Table 3.9, SigLIP models of different sizes show consistent accuracy improvements from fine-tuning on the Concept-pedia training data, demonstrating an enhanced ability to detect a broader range of concepts, with an average accuracy increase of 5.2 percentage points. For instance, SigLIP-

Semantic Type	SigLIP -base	SigLIP -base-ft	SigLIP -large	SigLIP -large-ft	SigLIP -SO400m	SigLIP -SO400m-ft
ANIM	59.6	60.4	66.9	67.1	69.2	71.9
BIO	51.1	55.1	58.4	62.0	60.1	63.0
FOOD	50.9	54.4	56.3	62.4	59.7	65.3
CEL	44.1	52.9	44.2	55.9	61.8	55.9
PLANT	40.1	46.0	45.3	53.7	49.1	55.3
MON	40.0	48.7	49.3	61.3	45.3	55.3
GROUP	34.1	37.8	40.6	43.7	44.3	48.2
ARTIFACT	32.4	39.8	36.5	45.2	41.3	48.2
DISEASE	30.7	30.7	35.3	36.6	42.4	46.0
PART	28.5	33.8	35.4	38.6	36.5	40.8
STRUCT	27.4	31.8	30.5	37.0	39.3	43.7
ASSET	25.9	35.8	27.2	44.4	33.3	46.9
EVE	24.1	28.7	28.2	34.4	33.4	36.1
LOC	23.7	26.5	28.9	31.7	32.6	33.4
DISCIPLINE	23.2	28.1	28.8	35.0	32.0	36.1
SUPER	23.1	31.3	31.3	34.3	31.3	45.5
SUB	22.3	28.3	28.0	34.5	33.3	38.4
MEDIA	20.6	25.9	25.5	32.0	28.7	34.3
PER	22.3	23.8	29.2	29.2	32.2	37.6
PROPERTY	22.1	26.6	27.2	30.9	30.3	30.3
LANGUAGE ITEM	17.9	22.9	22.9	27.1	27.9	30.0
CULT	20.0	21.0	22.1	26.7	26.7	30.3
ORG	18.8	19.8	22.7	27.1	30.5	35.9
RELATION	18.7	21.5	25.8	21.5	26.3	23.4
MEASURE	18.6	21.9	25.4	25.6	24.5	26.0
LAW	14.6	18.5	16.9	24.6	19.2	23.1
PSY	13.3	18.5	19.2	20.4	23.2	24.1

Table 3.10. We report the zero-shot and fine-tuned (ft) SigLIP models across all semantic categories presented in Concept-10k. Results in **bold** indicate the best performance per model size.

SO400m achieves 45.0% accuracy, up from 40.3%, highlighting that Concept-10k remains highly challenging even after fine-tuning. Moreover, we observe consistent performance improvements across semantic categories when models are fine-tuned using Concept-pedia, particularly in categories where models struggle in the zero-shot setting. As shown in Table 3.10, SigLIP-large shows a notable increase of 7.7 percentage points in the LAW category, while also improving on more common categories such as ANIM and FOOD. Interestingly, model performance on ImageNet-1k variations is minimally affected, despite differences in category coverage and distribution.

Training details. We fine-tuned open-sourced SigLIP-base ⁶, SigLIP-large ⁷, and SigLIP-SO400m ⁸ on the complete training split of Concept-pedia. The models are trained with a

⁶[google/siglip-base-patch16-256](https://github.com/google/siglip-base-patch16-256)

⁷[google/siglip-large-patch16-256](https://github.com/google/siglip-large-patch16-256)

⁸[google/siglip-so400m-patch14-384](https://github.com/google/siglip-so400m-patch14-384)

batch size of 256 for 50,000 steps using AdaFactor (Shazeer and Stern, 2018a) as optimizer with a learning rate of $1 \cdot 10^{-6}$.

Multi-Label Ablation We ablate the impact of leveraging multi-label annotations in Concept-pedia for fine-tuning. To do so, we construct a single-label version of Concept-pedia and fine-tune both SigLIP-base and SigLIP-large on it, comparing their performance against models trained using the original multi-label annotations. To generate the single-label version of Concept-pedia, we retain the label assigned by the propagation heuristic with the highest precision, as reported in Table 3.4. For samples with multiple annotations of the same type, we resolve conflicts by randomly selecting one label. As shown in Table 3.11, regardless of model size, while fine-tuning on single-label annotations still improves performance, models benefit even more from multi-label annotations. This confirms the advantage of using multi-label annotations for images when suitable and available.

Model	Zero-shot	Single-label	Multi-label
SigLIP-base	31.1	33.9	36.1
SigLIP-large	36.1	39.4	41.9

Table 3.11. Performance comparison between models fine-tuned on Concept-pedia using single-label and multi-label annotations.

3.7 Summary

Recent advances in vision-language modeling have been driven by large-scale pre-training and strong performance on established benchmarks such as ImageNet. At the same time, these benchmarks capture only a narrow subset of the semantic diversity encountered in real-world multimodal environments, limiting our ability to assess how broadly models represent visual concepts. Motivated by the fact that semantic concepts span a much broader and more varying granularity than what current evaluations reflect, this chapter introduced a scalable methodology for constructing wide-coverage multimodal resources grounded in encyclopedic knowledge.

We presented *Concept-pedia*, a large-scale semantically annotated multimodal resource automatically constructed from Wikipedia, together with *Concept-10k*, a manually

validated benchmark designed to evaluate visual concept recognition across a substantially broader range of semantic categories than existing datasets. Our empirical analysis showed that state-of-the-art VLMs, despite strong performance on standard benchmarks such as ImageNet-1k, struggle to generalize to the diverse and long-tail concepts included in Concept-10k. In the zero-shot setting, even SigLIP-SO400m, which achieves 83.2% top-1 accuracy on ImageNet, reaches only 40.3% on Concept-10k.

Furthermore, we demonstrate that Concept-pedia can be leveraged to improve model generalization to underrepresented concept categories. Across models, fine-tuning with Concept-pedia yields an average accuracy gain of 5.2 percentage points on previously unseen concepts, while preserving performance on established benchmarks. These results indicate that large-scale, semantically grounded resources can effectively expand multimodal knowledge coverage without sacrificing existing capabilities, and provide a practical framework for continuously extending visual concept recognition beyond narrowly scoped benchmarks.

More broadly, this chapter shows that progress in multimodal understanding requires moving beyond closed and narrowly scoped benchmarks toward resources that better reflect the breadth and diversity of real-world concepts. However, while expanding concept-level coverage is necessary for robust semantic grounding, many knowledge-intensive applications ultimately require identifying specific real-world instances rather than generic categories. This motivates the need to shift from conceptual abstraction toward entity-centric multimodal understanding, which is the focus of the next chapter.

Chapter 4

Large-Scale Multimodal Knowledge - Beyond Concepts

4.1 Overview

The previous chapter addressed limitations in multimodal knowledge coverage at the level of visual concepts, showing that current Vision-language Models (VLMs) struggle to generalize beyond the narrow semantic distributions captured by widely used benchmarks. While concept-level recognition provides an essential foundation for semantic understanding, many knowledge-intensive applications also require identifying and reasoning about specific real-world instances (i.e., entities).

Accordingly, this chapter shifts the focus from concepts to entities. In real-world settings, entity mentions often appear in multimodal contexts, making Multimodal Entity Linking (MEL) a natural testbed for studying entity-level multimodal knowledge access. Despite recent progress, existing MEL resources and models remain constrained by limited entity coverage, dataset-specific bias, and closed-world assumptions. Most benchmarks are derived from narrow subsets of Wikipedia, and models are typically trained and evaluated within the same data distribution, resulting in strong in-domain performance but limited robustness under distribution shift.

To address these limitations, this chapter introduces *Enti-pedia*, a large-scale multimodal resource derived from Concept-pedia annotations. Enti-pedia provides broad entity coverage together with substantial multimodal supervision per entity, enabling pre-training

of MEL models at scale. We further establish an experimental framework that evaluates both in-domain and cross-dataset performance across multiple MEL benchmarks, quantifying the extent of dataset overfitting and the benefits of large-scale pre-training for robust entity-level knowledge access.

Overall, this chapter extends the study of wide-coverage multimodal knowledge from concepts to entities, advancing the representation dimension of this dissertation toward more realistic, open-domain settings. The remaining limitation of static entity inventories motivates the next chapter, which investigates how to dynamically integrate emerging knowledge beyond fixed knowledge bases.

4.1.1 From Concepts to Entities

Concept recognition evaluates whether a model can associate multimodal evidence with semantic categories or properties (e.g., *dog*, *bridge*, *disease*). While this capability is essential for basic semantic grounding, it is insufficient for applications that require identifying specific real-world instances and linking them to structured repositories. Entities correspond to uniquely identifiable people, organizations, locations, events, or artifacts anchored in knowledge bases, and they serve as the referential units required by downstream tasks such as retrieval, reasoning, and knowledge integration.

Moving from concepts to entities introduces substantially higher demands on knowledge representation and disambiguation. Unlike concepts, which can overlap and admit multiple valid labels, entity identity is discrete: each mention must be resolved to a single referent. Moreover, disambiguating entities often requires fine-grained world knowledge and contextual cues, as many entities share identical surface forms or visually similar appearances. These challenges are further amplified in multimodal settings, where textual and visual signals may provide partially conflicting evidence, making entity resolution dependent on effective cross-modal reasoning.

Multimodal Entity Linking (MEL) formalizes this problem by requiring systems to leverage both textual context and visual information to resolve a mention to its correct knowledge base entry, serving as the primary task examined in this chapter.

4.2 Multimodal Entity Linking

Given a multimodal context consisting of text and associated images, MEL aims to link each entity mention to its corresponding entry in a knowledge base, typically Wikipedia. Formally, for a mention m occurring in textual context x and associated with an image I , the model predicts an entity $e \in \mathcal{E}$ from a predefined entity inventory $\mathcal{E}$:

$$\hat{e} = \arg \max_{e \in \mathcal{E}} s((m, x, I), e),$$

where $s(\cdot)$ scores the compatibility between the multimodal mention context and representations of candidate entities, which may include titles, descriptions, images, and other metadata. MEL can be instantiated under different candidate assumptions. In open-inventory settings, candidates may span the full knowledge base, requiring scalable retrieval and robust disambiguation. In many benchmark settings, candidate pools are restricted at the instance or dataset level, simplifying retrieval but potentially inflating performance and masking limitations in coverage and generalization.

4.2.1 Challenges and Limitations

Correct entity linking depends jointly on recognition and disambiguation under weak or incomplete evidence, requiring systems to integrate information across modalities while operating over large and highly ambiguous entity spaces. Beyond these inherent task-level difficulties, current MEL research exhibits two structural bottlenecks. First, existing datasets are typically constructed from narrow subsets of Wikipedia or specific domains, limiting entity diversity and biasing coverage toward frequent entities and recurring topics. Second, models are almost always trained and evaluated on the same dataset distribution, encouraging dataset-specific optimization and obscuring cross-dataset robustness.

As a consequence, strong in-domain accuracy often fails to translate into reliable performance under distribution shift, where entity distributions, visual styles, and contextual cues can differ substantially. These limitations constrain scalability and hinder the deployment of MEL systems in realistic open-domain environments, motivating the need for resources and evaluation protocols that explicitly target broad coverage and generalization.

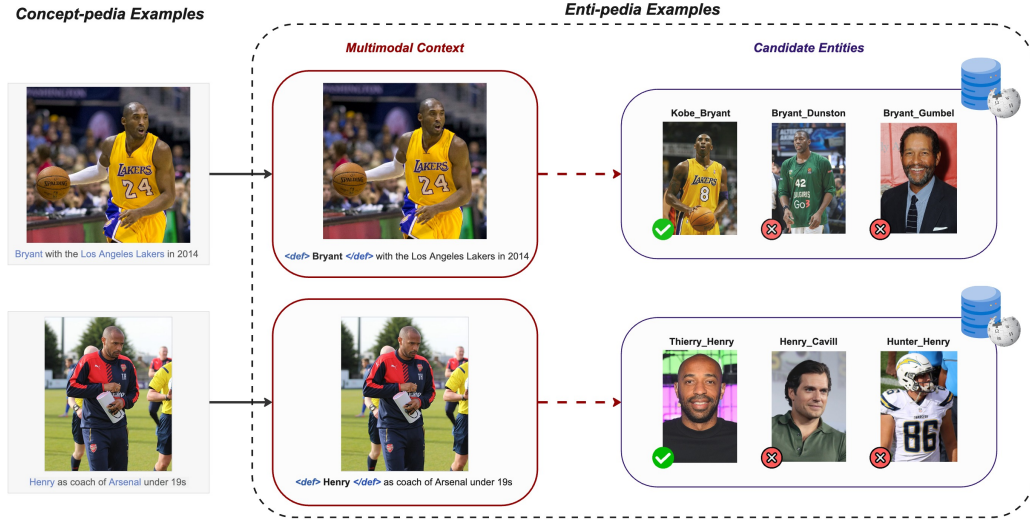


Figure 4.1. Examples for Enti-pedia creation procedure.

4.3 Enti-pedia: A Large-Scale Resource for Multimodal Entity Linking

To address the coverage and generalization limitations identified above, we introduce *Enti-pedia*, a large-scale multimodal resource designed to support entity-centric representation learning under open-domain coverage.

Wikipedia for EL across modalities. Our decision to ground Enti-pedia in Wikipedia is aligned with long-standing practice in both text-only Entity Linking (EL) and MEL. Classical EL benchmarks such as AIDA-CoNLL are built directly from English Wikipedia articles and links (Hoffart et al., 2011), and large-scale benchmarks for knowledge-intensive language tasks, such as KILT, explicitly standardize evaluation on a shared snapshot of Wikipedia as the underlying knowledge source (Petroni et al., 2021). Similarly, state-of-the-art EL systems like BLINK (Wu et al., 2020) and GENRE (De Cao et al., 2021) treat Wikipedia as the target entity repository, training models to retrieve or generate Wikipedia titles as canonical identifiers. This tight coupling between EL and Wikipedia naturally extends to multimodal settings: WikiMEL and WikiDiverse are both constructed from Wikipedia pages enriched with images, captions, and entity annotations (Yang et al., 2023; Wang et al., 2022c). Following this established practice, constructing Enti-pedia as a large-scale, Wikipedia-grounded multimodal resource ensures compatibility with existing EL and MEL

benchmarks while substantially expanding entity coverage and supervision density.

Construction. Enti-pedia is constructed by repurposing entity-level annotations already present in Concept-pedia, where each image is paired with a caption containing hyperlinks to Wikipedia entities. While Concept-pedia is used primarily for concept-level supervision, these hyperlink signals provide an additional layer of entity grounding that can be naturally reframed for MEL.

Concretely, for each image-caption pair, we extract mentions linked to Wikipedia and convert them into MEL training instances. Given an image I , caption context x , and a hyperlinked mention m linked to entity e , we construct training tuples of the form (m, x, I, e) . Following conventions introduced in autoregressive EL frameworks such as GENRE and adopted in later MEL work, we mark entity mentions with delimiter tokens to explicitly indicate the span to be resolved.

For example, consider an image of Kobe Bryant accompanied by the caption "*Bryant with the Los Angeles Lakers in 2014*", where "*Bryant*" is linked to the corresponding Wikipedia page. We rewrite the caption as:

`<def> Bryant </def> with the Los Angeles Lakers in 2014`

Here, the special tokens `<def>` and `</def>` identify the start and end of the entity mention to be linked, while the target label corresponds to the Wikipedia entry *Kobe_Bryant*. We provide additional examples from Enti-pedia and this conversion procedure in Figure 4.1. This transformation effectively converts Concept-pedia’s propagated annotations (see Figure 3.9) into explicit entity-centric supervision, enabling large-scale training of multimodal models that learn to resolve mentions provided within a multimodal (visual and textual) context.

Crucially, Enti-pedia inherits the broad topical and semantic diversity of Concept-pedia, which spans the entire English Wikipedia, and is therefore not restricted to a narrow subset of entities or a single benchmark distribution. Instead, it provides supervision for a large and diverse inventory of entities, each appearing across multiple multimodal contexts. By aggregating occurrences of the same entity across different documents and images, Enti-pedia offers substantially richer per-entity evidence than existing MEL datasets, which typically include only a handful of examples per entity.

Dataset	Entities	Pairs	Entities ≥ 5 ex.
WikiDiverse	$\approx 7\text{K}$	$\approx 8\text{K}$	≈ 200
WikiMEL	$\approx 13\text{K}$	$\approx 23\text{K}$	≈ 150
Enti-pedia	$\approx 1.2\text{M}$	$\approx 2.3\text{M}$	$\approx 200\text{K}$

Table 4.1. Scale comparison between existing MEL benchmarks and Enti-pedia. Enti-pedia provides orders of magnitude broader entity coverage and substantially more multimodal evidence per entity.

4.3.1 Enti-pedia Scale and Coverage

Table 4.1 quantifies the scale difference between Enti-pedia and commonly used MEL benchmarks. WikiDiverse contains 7,329 unique entities and 8,252 image-text pairs, with only 200 entities having more than five examples. WikiMEL expands this to 13,212 entities and 22,537 pairs, yet still provides dense supervision for just 150 entities. In contrast, Enti-pedia comprises approximately 1.2 million unique entities and 2.3 million multimodal instances, with over 200,000 entities supported by more than five examples. This increase of several orders of magnitude in both entity coverage and supervision density enables models to train over a substantially broader and more varied entity space, encouraging representations that generalize across datasets and domains rather than overfitting to the narrow distributions of individual benchmarks.

Wikipedia for large-scale EL pre-training. By enabling large-scale pre-training over a broad and diverse entity inventory, Enti-pedia provides the foundation for studying generalization in MEL beyond narrow benchmark supervision. This design mirrors the role that large-scale Wikipedia-derived training data has played in text-only EL, where resources such as the BLINK dataset (Wu et al., 2020), constructed by leveraging Wikipedia’s hyperlink structure as silver-standard entity annotations, have been instrumental in developing strong entity linking models like BLINK and GENRE (De Cao et al., 2021) that transfer robustly across benchmarks. Enti-pedia extends this strategy to the multimodal setting by similarly exploiting hyperlink annotations, in this case from Concept-pedia, to expose models to a wide range of entity types and multimodal contexts, encouraging representations that capture transferable entity knowledge rather than dataset-specific regularities. In the following sections, we describe the experimental setup used to evaluate the impact of Enti-

pedia pre-training and analyze its effect on both in-domain performance and cross-dataset robustness.

4.4 Experimental Setup

In this section, we describe the benchmarks, baselines, training procedure, and evaluation metrics used to assess large-scale MEL with and without pre-training on *Enti-pedia*.

4.4.1 Evaluation Benchmarks

To study entity-centric multimodal generalization, we evaluate our models using the two most widely used Wikipedia-grounded MEL benchmarks, namely, *WikiMEL* and *WikiDiverse*. Both datasets frame MEL as the task of linking entity mentions appearing in image-text contexts to their corresponding Wikipedia entries, making them suitable testbeds for assessing entity-level multimodal knowledge access.

WikiMEL (Yang et al., 2023) provides a large collection of image-text pairs with annotated entity mentions extracted from Wikipedia, though its entity distribution is notably skewed toward a small number of entity types, predominantly persons. WikiDiverse (Wang et al., 2022c), in contrast, is sourced from Wikinews and human-annotated across a diverse range of entity types, emphasizing topical and entity-type diversity to reduce the dominance of frequent categories.

Evaluating on both benchmarks allows us to assess model performance under a concentrated entity distribution as well as under greater entity-type and topical variation. Following prior MEL work, we use the official train, validation, and test splits reported in the original publications.

4.4.2 Models and Baselines

We compare our approach against representative state-of-the-art MEL systems that combine multimodal encoders with entity-aware matching. These models illustrate the dominant design paradigms in current MEL research, where architectures are typically optimized for benchmark-defined candidate sets.

Initial approaches include multimodal fusion architectures that learn joint image-text

representations for entity matching (Yang et al., 2023). RichpediaMEL enriches entity representations with structured and multimodal signals derived from Wikipedia pages, improving alignment between mention context and entity content (Wang et al., 2022b). GEMEL adopts a generative formulation that treats MEL as sequence generation, enabling a single framework to support multiple evaluation settings (Shi et al., 2024).

Additional representative models explore diverse strategies for cross-modal entity matching. MIMIC (Luo et al., 2023a) applies multi-grained interaction mechanisms to capture both intra-modal and inter-modal correspondences at local and global levels. OT-MEL (Zhang et al., 2024b) takes a different approach by formulating cross-modal correlation assignment as an optimal transport problem, replacing standard attention with globally informed transport plans. MELOV (Sui et al., 2024) instead focuses on the visual side, using variational autoencoders to generate text-conditioned visual features and optimize visual representations in latent space. On the other hand, M³EL (Hu et al., 2025) combines intra-modal and cross-modal matching at multiple granularities, including bidirectional text-visual interaction.

For all baselines, we report in-domain performance as provided in the original publications or reproduce results under the same benchmark-defined candidate sets and evaluation protocols used in WikiMEL and WikiDiverse.

4.4.3 Enti-pedia pre-training

To assess whether large-scale entity-centric supervision improves MEL beyond benchmark-specific training, we pre-train a MEL model on Enti-pedia and evaluate its transfer to standard benchmarks. The goal is to determine whether broad entity coverage and exposure to diverse multimodal contexts enable representations that generalize across datasets rather than overfit to a single benchmark distribution.

We adopt a dense bi-encoder formulation inspired by prior large-scale EL work (Wu et al., 2020), where multimodal mentions and entity representations are projected into a shared embedding space. FLAVA is selected as the backbone encoder because it is pretrained on multimodal data and produces unified image-text representations, making it well suited for handling the multimodal context provided in MEL. Training on Enti-pedia exposes the model to millions of entity-linked image-text instances spanning a long-tail

entity distribution, encouraging the model to associate mentions with entities under diverse visual and linguistic contexts.

We refer to the resulting pretrained model as *L-MEL* (Large-scale Multimodal Entity Linking), emphasizing that its representations are learned from wide-coverage entity supervision rather than benchmark-specific training. This mirrors the role of Concept-pedia at the concept level, where broad supervision improves generalization without compromising in-domain performance. As detailed in Section 4.5, L-MEL exhibits improved robustness under cross-dataset evaluation compared to systems trained only on benchmark data.

Training details. We pre-train the multimodal bi-encoder on the full Enti-pedia training corpus using contrastive learning with in-batch negatives. Each training instance consists of a multimodal mention context paired with its corresponding entity representation, and the objective maximizes similarity for correct pairs while pushing apart non-matching entities. Training is performed using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $1 \cdot 10^{-5}$, batch size 256, and linear warmup followed by cosine decay. Entity embeddings are periodically re-encoded to maintain consistency with encoder updates, and inference uses cosine similarity for nearest-neighbor retrieval.

4.4.4 Generalization Evaluation in EL

A central goal of Enti-pedia is to move MEL evaluation beyond benchmark-specific optimization toward assessing generalization under cross-dataset conditions. In practice, many MEL systems are trained and evaluated on the same dataset, which can inflate performance by allowing models to exploit dataset-specific regularities rather than learning transferable entity representations. As a result, strong in-domain accuracy does not necessarily reflect robustness in realistic settings, where entity distributions, visual styles, and contextual patterns differ from those observed during training.

This concern is well recognized in text-only EL. Although models are commonly trained on canonical datasets such as AIDA, their generalization capability is increasingly judged by performance on out-of-domain benchmarks rather than on the AIDA test set alone. Following the same principle, we adopt a cross-dataset evaluation protocol in addition to standard in-domain testing. Models are trained either on a benchmark dataset or on Enti-

pedia and then evaluated on multiple MEL benchmarks with differing entity distributions. This setup measures whether learned representations capture transferable entity semantics rather than exploiting benchmark-specific patterns.

All evaluations follow the standard MEL formulation, where a model resolves a multimodal mention to an entity drawn from a benchmark-defined candidate set. While these candidate pools do not fully approximate open-inventory linking over the entire Wikipedia entity space, cross-dataset transfer provides a practical diagnostic of robustness, enabling a clear distinction between systems optimized for a single dataset and systems that benefit from wide-coverage pre-training.

Evaluation Metrics. Following established MEL practice on WikiMEL and WikiDiverse, we use top-1 accuracy (Hits@1) as the primary evaluation metric (Yang et al., 2023; Wang et al., 2022c). Given a mention and its candidate set, a prediction is counted as correct if the entity with the highest similarity score matches the gold Wikipedia entity. To assess cross-dataset robustness, we additionally report the average of in-domain and out-of-domain scores when models are trained on one benchmark and evaluated on the other, highlighting the extent to which they generalize beyond their training distribution.

4.5 Results and Discussion

This section presents experimental results designed to evaluate both in-domain performance and cross-dataset robustness for MEL models, with and without wide-coverage pre-training on *Enti-pedia*.

4.5.1 In-domain versus Cross-dataset Evaluation

Table 4.2 summarizes Hits@1 accuracy when models are trained on either WikiMEL or WikiDiverse and evaluated both in-domain and across benchmarks. A consistent pattern emerges: models optimized for a specific benchmark often achieve competitive in-domain scores, but their performance degrades substantially when evaluated on a different dataset with a distinct entity distribution, as well as broader topical and entity-type coverage. This mirrors a broader concern in EL, where strong benchmark-specific performance does not necessarily translate to robust generalization across settings (Wu et al., 2020). For

completeness, Table 4.3 reports in-domain results from recent MEL systems that are trained and evaluated on a single benchmark without cross-dataset testing. While these results demonstrate the effectiveness of current multimodal interaction mechanisms, they do not necessarily indicate whether such performance reflects generalizable entity knowledge.

Entity inventories in current MEL systems. The two evaluation settings also differ in entity inventory scale. Models in Table 4.3 score each mention against benchmark-specific knowledge bases, while Table 4.2 evaluates against the full Wikipedia entity space. These benchmark inventories are considerably smaller than the millions of entries that text-only EL systems are typically expected to handle. As a result, even high in-domain accuracy under benchmark inventories does not directly indicate how a system would perform in a realistic linking scenario where the full Wikipedia entity space must be considered, and the gap with respect to open-inventory linking remains an open challenge for the field.

Model	Params	Training Data	WikiMEL	WikiDiverse	Avg.
GEMEL	7B	WikiMEL	63.57	71.49	67.53
GEMEL	7B	WikiDiverse	62.72	73.63	<u>68.18</u>
L-MEL	350M	Enti-pedia	<u>63.45</u>	<u>72.93</u>	68.19

Table 4.2. Cross-dataset transfer in MEL. Hits@1 scores on WikiMEL and WikiDiverse when models are trained on one benchmark and evaluated on both. **Avg.** is the macro-average over the two evaluation benchmarks. Results in **bold** indicate the best performance, while underlined values indicate the second best.

Model	WikiMEL	WikiDiverse	Avg.
MIMIC	87.98	63.51	75.75
OT-MEL	88.97	66.07	77.52
MELOV	<u>88.91</u>	67.32	78.12
M ³ EL	88.84	73.44	81.14
L-MEL	88.51	<u>68.03</u>	<u>78.27</u>

Table 4.3. Hits@1 scores on WikiMEL and WikiDiverse using benchmark-defined candidate sets. Baselines are trained on each benchmark’s training split. L-MEL on the other hand, is pretrained on Enti-pedia only, without fine-tuning on benchmark data. Baseline scores are reported from original publications. **Avg.** denotes the macro-average across the two benchmarks. Results in **bold** indicate the best performance, while underlined values indicate the second best.

4.5.2 Results

Cross-dataset MEL. Table 4.2 evaluates cross-dataset generalization under open-inventory evaluation conditions, where models retrieve from the full Wikipedia entity space. We include GEMEL in this comparison because its generative formulation supports constrained decoding over entity names, allowing evaluation against different target inventories by replacing the candidate entity set. We use the official GEMEL implementation and follow the OPT (Zhang et al., 2022a) configuration provided in the released codebase.

For each benchmark, GEMEL achieves its strongest results in-domain, but no single benchmark-trained configuration performs best across both benchmarks. When trained on WikiMEL, GEMEL attains 63.57 accuracy on WikiMEL and 71.49 on WikiDiverse. Conversely, when trained on WikiDiverse, it reaches 73.63 accuracy on WikiDiverse but only 62.72 on WikiMEL.

In contrast, L-MEL, pretrained on Enti-pedia without benchmark-specific training on either benchmark, achieves 63.45 on WikiMEL and 72.93 on WikiDiverse, yielding the highest macro-average (68.19). This result is particularly notable given that L-MEL is roughly an order of magnitude smaller than GEMEL (approximately 350M vs. 7B parameters) and, unlike GEMEL, does not benefit from in-context learning with retrieved instances. Although these absolute scores are lower than those reported in Table 4.3, where models are evaluated against smaller benchmark-specific inventories rather than the full Wikipedia entity space, the results suggest that wide-coverage entity-centric pre-training can produce competitive cross-dataset performance with a substantially simpler and more efficient architecture.

In-domain performance. Table 4.3 situates L-MEL within the broader landscape of benchmark-tuned MEL systems, where models are evaluated against benchmark-specific entity inventories. Among recent architectures, M³EL achieves the strongest macro-average (81.14), driven primarily by its WikiDiverse accuracy (73.44), while MIMIC, OT-MEL, and MELOV range from 75.75 to 78.12. Evaluated against the same benchmark inventories but without any fine-tuning on benchmark data, L-MEL achieves the second-strongest macro-average (78.27), surpassing MIMIC, OT-MEL, and MELOV, all of which are explicitly trained on each benchmark’s training split. Importantly, these results do not contradict the

transfer findings. Rather, they suggest that wide-coverage pre-training can approach, and in some cases surpass, specialized benchmark performance without benchmark-specific training. It is worth noting that we were unable to include these models in the cross-dataset evaluation of Table 4.2, as none of them release trained model checkpoints. The scores reported in Table 4.3 are therefore taken directly from their respective publications.

4.5.3 Discussion

The contrast between Tables 4.2 and 4.3 highlights a key finding: wide-coverage pre-training can improve cross-dataset robustness without sacrificing in-domain accuracy. L-MEL matches, and sometimes exceeds, benchmark-tuned systems in their own evaluation setting while also generalizing across datasets, suggesting that the two objectives can be jointly addressed when the model is exposed to a sufficiently broad entity distribution during pre-training. Furthermore, we show that this can be achieved using a comparatively lightweight bi-encoder architecture, without the complex interaction modules prevalent in recent MEL systems. This is consistent with broader evidence in text-only EL that broad entity coverage and domain-agnostic training objectives support stronger generalization beyond a single benchmark setting (Wu et al., 2020).

The cross-dataset results in Table 4.2 point to a deeper issue: current MEL benchmarks, while valuable for controlled comparison, may reward dataset-specific regularities that do not transfer. WikiMEL, for example, is skewed toward person entities, whereas WikiDiverse was explicitly designed to provide broader topical and entity-type diversity. A model that adapts closely to the dominant patterns of one benchmark may perform well in-domain but lack the representational breadth needed for the other. Enti-pedia mitigates this by exposing the model to a wider distribution of entity types and multimodal contexts during pre-training, reducing reliance on the idiosyncrasies of any single benchmark.

At the same time, important challenges remain. Entity distributions are inherently long-tailed, and our results do not isolate whether L-MEL’s gains extend to rare entities that appear infrequently even in Enti-pedia. Furthermore, multimodal evidence may be incomplete or only weakly aligned with the mention context, a limitation that affects current MEL systems and that benchmark evaluations do not explicitly measure. Addressing these issues requires future work on rare-entity representation, robustness to missing or noisy

modalities, and evaluation under open-inventory conditions across diverse domains and entity types that more closely approximate real-world scenarios.

Overall, these results reinforce the central narrative of this dissertation: progress toward knowledge-intensive multimodal capabilities requires not only improved architectures and benchmarks, but also resources that reflect the breadth and diversity of real-world entities.

4.6 Summary

This chapter extends the study of multimodal knowledge representation from concept-level coverage to entity-centric knowledge access through the task of MEL. While recognizing broad semantic categories is necessary for multimodal understanding, many knowledge-intensive applications require resolving mentions to specific real-world entities grounded in structured inventories. We formalize MEL as a setting in which a system must disambiguate entity mentions in multimodal contexts, and show that benchmark-centric evaluation can obscure a key limitation: strong in-domain performance does not reliably indicate robustness across datasets.

To address this limitation, we introduce *Enti-pedia*, a large-scale multimodal pre-training resource constructed from Wikipedia-grounded entity mentions aggregated across diverse contexts. Compared to existing MEL benchmarks, Enti-pedia provides substantially broader entity coverage and denser supervision per entity, enabling models to learn representations that better capture transferable entity-level knowledge. Empirical results show that pre-training on Enti-pedia improves cross-dataset generalization under open-inventory evaluation. Whereas GEMEL’s performance varies depending on which benchmark it is trained on, the Enti-pedia-pretrained L-MEL model achieves more balanced performance across WikiMEL and WikiDiverse, yielding the highest macro-average (68.19) with roughly an order of magnitude fewer parameters.

At the same time, entity-centric multimodal linking remains challenged by long-tailed entity distributions and imperfect multimodal alignment, where relevant visual or textual evidence may be incomplete or weakly coupled to the mention context. These limitations highlight the need for approaches that better capture rare entities, reconcile missing modali-

ties, and move toward open-inventory linking conditions that better approximate real-world settings.

Overall, these results show that broad entity coverage is important for building multimodal systems capable of reliable entity-level knowledge access, and that cross-dataset evaluation is necessary to assess whether such access genuinely generalizes beyond individual benchmarks. However, even with improved static representations, current methods remain tied to fixed entity inventories and are not designed to handle the continuous emergence of new entities. Real-world knowledge continues to evolve as new entities appear and existing information changes, a setting that fixed pre-training resources such as *Enti-pedia* cannot address on their own. This motivates the next chapter, which investigates retrieval-augmented methods for dynamically incorporating emerging knowledge beyond fixed entity inventories.

Chapter 5

Emerging Knowledge Augmentation

5.1 Overview

The previous chapters introduced large-scale multimodal resources that substantially expanded the coverage of concepts and entities, addressing key limitations in how current models represent world knowledge. By broadening the semantic scope of training and evaluation, these resources also improved generalization beyond narrow benchmark distributions. However, even comprehensive resources remain inherently static: they reflect a snapshot of the world at a particular point in time and assume that the set of relevant entities is fixed at training and inference time. This static-world assumption is deeply embedded in traditional Entity Linking (EL), where systems typically operate over a predefined inventory of candidates derived from a fixed knowledge base snapshot (most commonly Wikipedia). This means that when an entity is absent from this inventory, the system has no corresponding entry to link to and, as a consequence, cannot represent or reason about that entity within the downstream pipeline.

In practice, this limitation becomes especially problematic in dynamic information environments. The real world evolves continuously: entities change attributes, roles, and relevance over time, and entirely new entities emerge as new public figures rise, organizations are founded, products are released, and scientific developments are introduced. Because updating knowledge repositories through manual curation or periodic reannotation is costly and slow, static inventories inevitably lag behind the evolving world. For example, a newly released product, or a recently appointed public official may be mentioned widely

in news and social media before appearing in Wikipedia or becoming reliably linkable under a standard EL setup. As a result, knowledge-intensive systems that depend on static inventories can fail precisely when up-to-date knowledge is most needed.

These constraints motivate methods that can access up-to-date information at inference time, rather than relying solely on static inventories or parametric knowledge. Retrieval-augmented modeling has therefore gained traction in knowledge-intensive NLP: by retrieving relevant evidence from large and continuously updated corpora, models can ground their predictions in external sources and reduce dependence on information implicitly stored in model parameters. This capability is especially valuable under emerging knowledge, where retrieval can surface contemporaneous evidence even when an entity is absent from the target inventory, enabling the system to construct a usable entity representation from available textual context.

In this chapter, we address the emerging-entity challenge by introducing *RAED* (Retrieval-Augmented Entity Descriptions), a framework for generating grounded entity descriptions from retrieved real-world information rather than relying on predefined candidate pages. RAED enables the creation of entries for entities that are absent from static inventories, and supports downstream use by allowing these generated descriptions to be incorporated into existing inventories over time.

More specifically, the contributions of this chapter are as follows:

- We introduce RAED, a retrieval-augmented framework that generates entity descriptions (titles and definitions) for emerging entities, enabling effective disambiguation even in the absence of a predefined entity page in the inventory.
- We propose a retriever trained to retrieve diverse, contextually relevant passages, shifting the focus from retrieving a single static summary to retrieving diverse, relevant evidence.
- We show that descriptions generated by RAED can be used to extend previously static inventories and be integrated into existing downstream components, supporting Emerging Entity Linking (EEL) as the world changes.

Overall, this chapter complements the static coverage advances of the previous

chapters with a mechanism for dynamic knowledge access, bridging the gap between wide-coverage representation and the continuously changing nature of real-world knowledge.

5.2 Static Knowledge and Emerging Entities

As previously mentioned, knowledge-intensive NLP systems are typically developed under the assumption that the set of relevant entities and facts is fixed at training and test time. In practice, however, world knowledge evolves continuously. New entities such as public figures, organizations, scientific discoveries, and world events emerge at a rapid pace, while attributes and facts associated with existing entities continue to evolve. This temporal dimension of knowledge introduces challenges that static resources and closed-world modeling assumptions cannot adequately address. Traditional Entity Linking (EL) and Entity Disambiguation (ED) systems rely on predefined knowledge bases, most commonly Wikipedia, and assume that each target entity is associated with a title and a definition (together referred to as an entity description) that uniquely identify it (Wu et al., 2020; De Cao et al., 2021; Procopio et al., 2023). This assumption implicitly treats the knowledge base as complete and up to date; when descriptions are missing, outdated, or incomplete, performance degrades sharply, particularly in low-resource languages and rapidly evolving domains.

However, this assumption breaks down for emerging or poorly described entities, as well as in lower-resource languages, limiting the robustness and scalability of existing systems. For example, consider querying a system trained before the Covid-19 pandemic with the sentence: *By this point, most people have had at least one brush with Covid-19.* Since *Covid-19* is an emerging entity absent from the training data, the model fails to correctly resolve the mention. Similarly, candidate retrieval-based approaches (Zhang et al., 2022b; Wang et al., 2024; Orlando et al., 2024) fail when they rely on outdated or incomplete entity inventories. This is true even in cases where the context is multimodal, as with MEL, if the entry is missing from the knowledge-base there is simply no way to correctly link the mention. Consider our example again: if an entry for *Covid-19* is not available in the knowledge base, the model is forced to select from incorrect candidates. Instead, if we retrieve general textual knowledge, such as *Covid-19 being shorthand for ‘coronavirus disease 2019’*, and condition the model on this relevant context, we enable it to generate

more accurate and informative outputs, even for previously unseen entities.

5.2.1 Challenges and Limitations

Retrieval-Augmented Information Extraction. Recent work has increasingly turned to retrieval-based approaches as a mechanism for coping with the dynamic nature of knowledge. Retrieval-Augmented Generation enables models to access external corpora at inference time, reducing dependence on parametric knowledge encoded during pre-training (Lewis et al., 2020b; Izacard and Grave, 2021; Zhang et al., 2022b). By conditioning generation on retrieved evidence, these methods provide a principled way to incorporate newly available information without requiring costly retraining. Retrieval-based architectures have significantly advanced Information Extraction (IE) tasks by incorporating external knowledge at inference time. Dense retrieval models, such as EntQA (Zhang et al., 2022b) and ReLiK (Orlando et al., 2024), have demonstrated improvements in linking mentions to entities by retrieving high-quality candidates from large-scale indices. Our approach also builds on the retrieval-augmented paradigm; however, instead of retrieving target candidate entities, it leverages general-purpose passages. This enables not only disambiguation, but also the generation of new entity titles and definitions. As a result, it offers a more flexible and scalable alternative to traditional methods that rely on predefined entity inventories. This also opens the door to automatically updating entity inventories, which frequently become outdated and are costly to maintain manually.

Generative Approaches for Entity Disambiguation. Autoregressive models have emerged as strong alternatives to classification-based ED systems. GENRE (De Cao et al., 2021) formulates ED as a sequence-to-sequence task, generating existing entity titles directly.

More recently, FusionED (Wang et al., 2024) employs an encoder-decoder model that uses candidate entity descriptions retrieved from an inventory. The encoder models interactions between the context and each candidate, producing individual representations, while the decoder selects the best match by generating its index from the ordered candidate list. Our approach builds on these insights by learning to generate entity descriptions. However, unlike FusionED, which narrows the search space by retrieving candidate entities directly, RAED enriches its input with passages retrieved from broad informative text. This

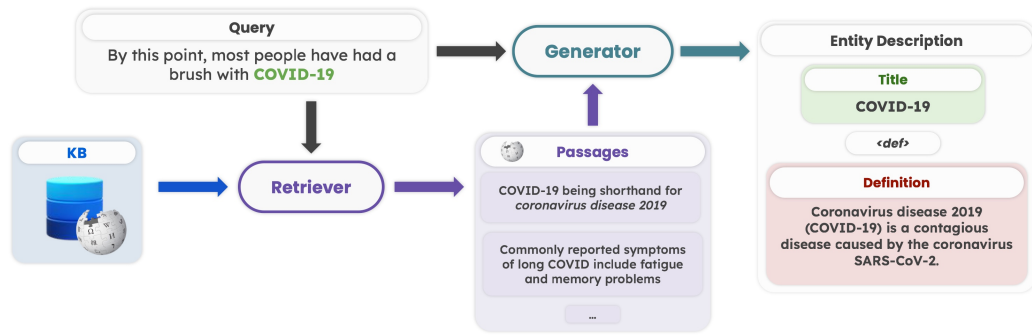


Figure 5.1. The Retrieval-Augmented Entity Descriptions (RAED) pipeline.

enables it to handle emerging entities that are absent from the inventory, a capability current models lack.

Handling Emerging Entities. A core challenge in ED and EL is handling new entities that are missing from the knowledge base at training time, known as emerging entities. Traditional systems assume a static knowledge base, making them ineffective for newly introduced entities. To study this issue, TempEL (Zaporojets et al., 2022) introduced a temporally segmented benchmark that evaluates the ability of models to adapt to new entities over time. EDIN (Kassner et al., 2022) extended this by proposing an end-to-end framework and benchmark to discover and index unknown entities, highlighting the importance of dynamically updating entity representations. To the best of our knowledge, the EDIN benchmark has not been publicly released. Despite these contributions, existing approaches do not address the challenge of automatically creating entries for missing entities.

Retrieval-Augmented Entity Description. To address these persistent issues, we propose RAED, a retrieval-based framework designed specifically to tackle the emerging entity challenge. Given a single mention of an entity, RAED generates a title and a definition by retrieving high-quality, contextually relevant information from external sources such as Wikipedia. Unlike traditional approaches, which are restricted to pre-constructed entity inventories (i.e., knowledge bases), our method focuses on generating meaningful descriptions, even for entities absent from the inventory.

5.3 RAED

In this section, we introduce RAED, our model for Entity Description Generation. RAED retrieves highly informative passages from a knowledge base (i.e., Wikipedia) and provides them as additional context to the generator to ensure that:

1. Generated descriptions are grounded in factual knowledge, thus limiting hallucinations.
2. Ambiguous entities are more effectively disambiguated, particularly when the original context alone is insufficient.

5.3.1 Passage Index

As discussed, our goal is to construct an external index that functions as a non-parametric memory. Following prior work (Lewis et al., 2020b; Izacard and Grave, 2021), we build this index using Wikipedia¹ by extracting the text from each page along with its *wikilinks*, i.e., hyperlinks to other Wikipedia pages. Previous studies have used wikilinks as pseudo-labels, treating them as disambiguation signals for linked spans (Wu et al., 2020; De Cao et al., 2021; Huguet Cabot and Navigli, 2021), while Izacard et al. (2021) leveraged entire entity pages as retrieval targets for training. In contrast, we leverage wikilinks as retrieval targets by retrieving passages where an entity is explicitly mentioned, rather than relying solely on its main Wikipedia page or description. In EL, retrieving entity descriptions is a high-precision objective (Zhang et al., 2022b; Orlando et al., 2024; Wang et al., 2024), but disambiguation fails when descriptions are missing or not retrieved. By shifting from descriptions to passages that provide additional context, i.e., either by directly mentioning the entity or containing relevant information, we enable the generator to produce more meaningful descriptions, even for entities not present in the inventory.

To construct our index, we follow these steps:

Wikipedia Parsing We use two libraries, *wikiextractor* (Attardi, 2015) and *mwparser-fromhell* (Kurtovic, 2013), to extract raw text from Wikipedia pages, including wikilinks

¹English Wikipedia dump released on Jan 20, 2023.

and boundary annotations. Disambiguation pages and other non-informative content are filtered out, yielding approximately 6 million extracted pages.

Entity Filtering To ensure a manageable index size, we divide each page into non-overlapping 100-word windows and retain only those that contain entities present in our datasets, either as annotated or candidate entities. This step results in a corpus of 88 million windows.

Similarity Filtering To further refine the index and retain meaningful content, we limit the number of passages per entity to 10. Passages are ranked by their similarity to the entity’s definition using a sentence similarity model,² and only the top-ranked ones are retained. The resulting index comprises 9 million passages.

It is important to note that while wikilinks are used during index construction, the retriever operates only on the raw text, without any direct access to the wikilinks or the associated entities.

5.3.2 Retriever

For the Retriever component, we follow a retrieval paradigm similar to that of DPR (Karpukhin et al., 2020, Dense Passage Retrieval), which uses an encoder to produce dense representations of queries and passages. In our setup, the input text q contains the target entity e for which we want to retrieve relevant passages. We use special tokens, i.e., [DEFINE] and [/DEFINE], to mark the beginning and end, respectively, of the entity mention. For example, the query *most people have had at least one brush with Covid-19* is represented as `most people have had at least one brush with [DEFINE] Covid-19 [/DEFINE]`.

Given $q \in \mathcal{Q}$ as our query in a collection of queries $\mathcal{Q}$ and a passage $p \in \mathcal{P}$ in a collection of passages $\mathcal{P}$ that corresponds to the Wikipedia passages or windows, the Retriever model computes:

$$E_Q(q) = \text{Retriever}(q), E_P(p) = \text{Retriever}(p)$$

²<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

and ranks the most relevant passages with respect to q using the similarity function $\text{sim}(q, p) = E_Q(q)^\top E_P(p)$, where the contextualized hidden representation of a query q and a passage p are the average of the encodings for the tokens in each of the two sequences computed by the same $\text{Retriever}(\cdot)$ Transformer encoder.

We train the Retriever employing a contrastive cross-entropy loss similar to SimCSE (Gao et al., 2021) as a training objective. The $\mathcal{L}_{\text{Retriever}}$ loss for q is defined as:

$$-\sum_{p^+ \in \mathcal{P}^+(q)} \log \frac{e^{\text{sim}(q, p^+)}}{e^{\text{sim}(q, p^+)} + \sum_{p^- \in \mathcal{P}^-(q)} e^{\text{sim}(q, p^-)}} \quad (5.1)$$

where $\mathcal{P}^+(q)$ are the gold passages of the entity present in q , and $\mathcal{P}^-(q)$ is the set of negative examples for q constructed using in-batch negatives from gold passages of other queries.

5.3.3 Entity Description Generation

We formalize the entity description task as follows: given a mention m in a query q and a set of passages $\mathcal{P}(q)$ relevant to q , RAED aims to generate a target description d for the corresponding entity:

$$P(d \mid q, \mathcal{P}(q)) = \prod_{k=1}^{|d|} P(d_k \mid d_{0:k-1}, q, \mathcal{P}(q))$$

Here, q refers to the same query introduced in Section 5.3.2, used as input for retrieving relevant passages. Specifically, we first retrieve the most relevant passages $\mathcal{P}(q)$, and then use those passages, along with the query q , to condition the generation of the entity description. The target description d consists of the Wikipedia title and the article’s opening sentence, separated by a special token $\langle \text{def} \rangle$, as shown in Figure 5.1.

5.3.4 RAED for Entity Disambiguation

Given that our objective is to generate a meaningful description for a given entity mention, our system indirectly learns to disambiguate that entity. Hence, here we show that RAED, although trained against a generative objective, can also be applied to discriminative tasks such as ED without the need for additional training. ED aims to identify the correct entity from a set of candidates $\mathcal{E}(m)$ for a given mention m . To directly utilize the descriptions

generated by RAED for ED, we employ sentence-transformers³ to compute a similarity score between the generated description $\hat{d}$ and a candidate reference description d_e of the entity e . We first encode both the generated description and all the candidates provided for m into vector representations (i.e., embeddings). We then select the candidate with the highest cosine similarity to the generated description, as follows:

$$\hat{e} = \arg \max_{e \in \mathcal{E}(m)} \text{sim}(\hat{d}, d_e) \quad (5.2)$$

5.4 Experimental Setup

In this section, we outline the experimental setup for training and evaluating RAED. Specifically, we describe the training procedures for the retriever and generator components (Sections 5.4.1 and 5.4.2), the datasets used (Section 5.4.3), and the evaluation metrics used (Section 5.4.4).

5.4.1 Retriever

We train the E5⁴ (Wang et al., 2022a) encoder directly on our Passage Index described in Section 5.3.1, using a self-supervised approach similar to Gao et al. (2021). Specifically, we treat the collection of passages $\mathcal{P}$ as our set of queries $\mathcal{Q}$. For each query $q \in \mathcal{Q}$, we consider as positive passages those $p \in \mathcal{P}^+(q)$ that refer to the same entity e as q . Negative passages are sampled from passages within the same batch that do not mention the same entity. We optimize the Noise Contrastive Estimation (NCE) loss (see Eq. 5.1 in Section 5.3.2) using 400 negatives per batch. The encoder is trained for a maximum of 100,000 steps using RAdam (Liu et al., 2020) with a learning rate of $2 \cdot 10^{-5}$ and a linear learning rate decay schedule.

5.4.2 Entity Description Generation

Our experiments explore various text generation models for the entity description generation task. Specifically, we evaluate both encoder-decoder models, including T5-large⁵ (Raffel

³<https://huggingface.co/Alibaba-NLP/gte-modernbert-base>

⁴<https://huggingface.co/intfloat/e5-base>

⁵<https://huggingface.co/google-t5/t5-large>

et al., 2020) and Flan-T5-large⁶ (Chung et al., 2024), and decoder-only models, such as SmolLM2-360M⁷ (Allal et al., 2025) and Llama-3.2-1B⁸ (Dubey et al., 2024). For each input text in AIDA (referred to here as the query q), we combine q with its top 10 retrieved passages to create the input for RAED. We fine-tune each model on the AIDA training split to generate a description for the mention m within q for 100,000 steps with a batch size of 32 input texts, using the AdamW optimizer (Loshchilov and Hutter, 2019) and a learning rate of $1 \cdot 10^{-6}$. To assess the impact of retrieved passages, we train each model to generate descriptions using the same data and setup, once with retrieval (denoted as $\text{Model}_{\text{RAED}}$) and once without retrieval (denoted as Model).

Integration of Retrieved Passages For encoder-decoder models, we integrate retrieved passages using the Fusion-in-Decoder (FiD) approach (Izacard and Grave, 2021), in which each retrieved passage is concatenated with the query, independently encoded, and then passed to the decoder. For decoder-only models, we incorporate the retrieved passages directly into the prompt.

5.4.3 Datasets

Wikipedia As explained in the previous section, we leverage the entire English Wikipedia to construct our index, which is used both to train the retriever and as the source of additional context in our RAED framework.

AIDA CoNLL-YAGO We adopt the experimental setup of De Cao et al. (2021), utilizing the standard AIDA-CoNLL splits (Hoffart et al., 2011) for training (AIDA-train), model selection (AIDA-testa), and in-domain evaluation (AIDA-testb). AIDA is a widely used benchmark for ED and EL, allowing consistent comparison with prior work and ensuring a robust evaluation framework.

Out-of-Domain (OOD) Datasets For out-of-domain ED evaluation, we use MSNBC (Cucerzan, 2007), AQUAINT (Milne and Witten, 2008), ACE2004 (Ratinov et al., 2011), WNED-CWEB (CWEB), and WNED-WIKI (WIKI), as curated by Alani et al. (2018)

⁶<https://huggingface.co/google/flan-t5-large>

⁷<https://huggingface.co/HuggingFaceTB/SmolLM2-360M>

⁸<https://huggingface.co/meta-llama/Llama-3.2-1B>

Dataset	Model	Retrieval	BL	R-1	BS	SIM	F-NLI
$AIDA_{testb}$	T5	$\times$	58.0	73.5	92.6	93.2	22.1
	T5 _{RAED}	FiD	70.9	82.7	96.2	95.3	49.9
	Flan-T5	$\times$	60.5	77.5	93.1	92.9	24.8
	Flan-T5 _{RAED}	FiD	70.6	82.5	96.2	95.2	50.2
	SmolLM2	$\times$	61.2	74.1	94.3	92.8	27.4
	SmolLM2 _{RAED}	Prompt	72.1	83.1	95.6	94.3	52.9
	Llama-3.2	$\times$	63.2	76.8	95.8	93.7	33.8
	Llama-3.2 _{RAED}	Prompt	73.9	84.0	96.9	95.4	55.5
OOD	T5	$\times$	21.0	45.8	88.1	81.2	-8.2
	T5 _{RAED}	FiD	45.1	63.8	92.4	87.8	25.6
	Flan-T5	$\times$	22.9	48.1	89.5	81.9	-3.7
	Flan-T5 _{RAED}	FiD	45.9	64.4	92.7	88.1	25.9
	SmolLM2	$\times$	24.4	47.0	88.7	83.0	-5.5
	SmolLM2 _{RAED}	Prompt	42.8	61.5	91.5	85.1	21.3
	Llama-3.2	$\times$	28.4	51.6	90.3	84.5	-2.0
	Llama-3.2 _{RAED}	Prompt	45.1	63.3	92.6	87.5	27.5

Table 5.1. Evaluation results of various models on the AIDA test split ($AIDA_{testb}$) and out-of-domain (OOD) datasets (MSNBC, AQUAINT, ACE2004, CWEB, and WIKI). Metrics include BLEU (BL), ROUGE-1 (R-1), BERTScore (BS), Semantic Similarity (SIM), and Factual-NLI (F-NLI). Models with $\times$ do not use retrieved passages. **Bold** indicates the best performance. All reported models were fine-tuned on the AIDA training set.

and Gabrilovich et al. (2013). These datasets span a diverse range of domains and text styles, providing a challenging benchmark for assessing model generalization. To ensure consistency with prior work, we use the same candidate sets originally introduced by Le and Titov (2018).

TempEL We also use TempEL (Zaporojets et al., 2022) which provides ten yearly snapshots of English Wikipedia entities, enabling the study of temporal dynamics in EL. These snapshots allow us to simulate the introduction of new entities into the knowledge base and to assess our models’ ability to generate descriptions for entities not seen at training time. TempEL also provides two types of entities: *continual entities*, which appear in all TempEL temporal snapshots, and *emerging entities*, which are newly introduced in specific snapshots and serve as the focus of our experiments.

5.4.4 Evaluation Metrics

We evaluate the quality of the generated entity descriptions using the following metrics:

Natural Language Generation (NLG) We employ traditional metrics like BLEU and ROUGE (Papineni et al., 2002; Lin, 2004) which measure n-gram overlap, capturing lexical similarity. However, these metrics rely on exact string matches and may not fully reflect the quality of the generated text. To address this limitation, we assess the Semantic Similarity (SIM)⁹ by calculating the cosine similarity between sentence embeddings of the generated and reference descriptions, focusing on semantic alignment. Additionally, we utilize BERTScore (Zhang et al., 2019), which evaluates token-level semantic similarity using contextualized embeddings from a pre-trained BERT model. By incorporating these complementary metrics, we aim to provide a comprehensive evaluation of the quality of the generated descriptions.

Natural Language Inference (NLI) Inspired by prior work (Chen and Eger, 2023; Scirè et al., 2024b), we evaluate the logical relationship between the generated and reference descriptions using a pre-trained NLI model.¹⁰ Specifically, we compute the difference between the probabilities assigned to the entailment and contradiction classes to assess factual consistency. We refer to this metric as Factual-NLI (F-NLI).

InKB Micro F1 This metric, specific to Entity Disambiguation (ED), computes the micro-averaged F1 score over mentions correctly linked to entities present in the knowledge base (InKB).

5.5 Results

In this section, we present a three-fold evaluation of RAED. First, we assess the quality and factual accuracy of the generated entity descriptions. Next, we evaluate their effectiveness in downstream tasks, including ED and EEL. Finally, we conduct a qualitative evaluation using an LLM-as-a-Judge framework to compare RAED’s descriptions against descriptions generated without retrieval.

⁹We use the same model as introduced in Section 5.3.4.

¹⁰<https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli>

Model	Retrieval	In-domain	Out-of-domain					AVG	
		AIDA-B	MSNBC	AQUAINT	ACE2004	CWEB	WIKI	ALL	OOD
GENRE	-	88.6	88.1	77.1	82.3	71.9	71.7	80.0	78.2
FusionED	<i>Candidates</i>	<u>91.7</u>	<u>92.4</u>	82.0	87.1	75.8	78.6	84.6	83.2
T5	-	90.1	91.3	78.4	89.4	74.4	72.2	82.6	81.1
T5 _{RAED}	<i>Passages</i>	<u>91.7</u>	92.3	82.0	<u>90.7</u>	<u>77.3</u>	<u>78.7</u>	85.5	84.2
FLAN-T5	-	89.4	91.5	79.5	91.4	74.7	73.4	83.3	82.1
FLAN-T5 _{RAED}	<i>Passages</i>	91.8	93.2	80.3	90.3	77.1	78.1	<u>85.1</u>	<u>83.8</u>
SmolLM2	-	89.1	91.6	82.3	85.8	76.3	77.5	83.8	82.7
SmolLM2 _{RAED}	<i>Passages</i>	91.0	91.9	<u>83.6</u>	88.3	76.0	77.1	84.5	83.4
Llama-3.2	-	91.0	92.1	82.1	<u>90.7</u>	74.9	75.0	84.3	83.0
Llama-3.2 _{RAED}	<i>Passages</i>	<u>91.7</u>	92.3	84.3	<u>89.5</u>	78.1	79.0	85.0	83.4

Table 5.2. InKB Micro F1 comparison of different models fine-tuned on the AIDA training set and evaluated on both in-domain and out-of-domain datasets. Results for GENRE and FusionED are reported as presented in their respective publications. **Bold** values indicate the best performance, while underlined values denote the second best.

5.5.1 Entity Description Generation

In Table 5.1, we present the results of the Entity Description Generation task across various configurations of RAED. Our retrieval strategy consistently improves performance across all models and sizes, as evidenced by increases in all NLG metrics. This indicates that the descriptions generated by RAED are more lexically and semantically aligned with the gold descriptions. Notably, we observe an average increase of 27.5 points in Factual-NLI scores across RAED models, suggesting that descriptions generated by RAED are significantly more factually consistent with the reference descriptions. We attribute these improvements to RAED’s grounded generation process, which reduces reliance on parametric memory, mitigates hallucinations, and leads to improved factual consistency. Similar gains are observed across all metrics for out-of-domain (OOD) datasets, highlighting RAED’s capability to generate semantically and factually accurate descriptions for entities unseen during training. For decoder-only models, an example of a prompt used to generate the entity description (title and definition) is shown in Table 5.3.

Efficiency RAED provides not only performance gains but also efficiency benefits. It outperforms both same-sized and significantly larger models that do not use retrieval. For instance, RAED, with SmolLM2 (360M parameters) as the generator and E5-base (109M parameters) as the retriever (totaling approximately 469M parameters), outperforms Llama-3.2 (1B parameters) across all metrics and achieves a 23.3-point gain in factuality, despite

Generate a title and definition separated by <def> for the mention of an entity between [DEFINE] and [/DEFINE]. Input Text: INPUT_TEXT Passages: Passage 1 ... Passage 2 ... Passage 3 ... Passage 4 ... Passage 5 ... Target Entity Description: “COVID-19 is a contagious disease caused by the coronavirus SARS-CoV-2”.

Table 5.3. Prompt for Entity Description Generation by RAED.

being less than half its size. This result demonstrates that incorporating external knowledge while reducing reliance on parametric memory can improve both performance and efficiency.

5.5.2 Entity Disambiguation

Table 5.2 presents our results using RAED as a discriminator for ED, as detailed in Section 5.3.4. Compared to models without retrieval, RAED variants show improvements across nearly all benchmarks, in both in-domain and out-of-domain settings. The most notable gains are observed with $T5_{RAED}$ and FLAN- $T5_{RAED}$, which outperform T5 and FLAN-T5 by 2.9 and 1.8 points, respectively.

Remarkably, RAED models not only match but also slightly outperform systems like FusionED, which are specifically trained for ED and provided with candidate entities as input. For example, although it uses the same underlying architecture (FLAN-T5-large), FLAN- $T5_{RAED}$ outperforms FusionED on most out-of-domain datasets, achieving an average improvement of 0.6 points, while also slightly surpassing it on the in-domain benchmark (91.8 vs. 91.7). This result is particularly noteworthy because FusionED is explicitly trained to select from a set of candidate entities, whereas RAED relies solely on retrieved passages to generate descriptions and learns to disambiguate mentions implicitly, without any direct ED supervision during training. Furthermore, our approach enables models like $T5_{RAED}$ to outperform larger and more recent models such as Llama-3.2-1B, achieving a 1.2-point overall improvement and a 1.2-point gain in the out-of-domain setting.

These results, obtained without access to candidate entities at inference time, show

that RAED exhibits strong generalizability and versatility. This highlights a key advantage of our approach: RAED can generate meaningful and discriminative descriptions using retrieved passages, eliminating dependence on predefined candidates even when the correct entity is absent from the knowledge base.

Model	Desc	2020	2021	2022
-	Gold	71.0	64.8	72.3
2019	T5	51.2	37.9	44.5
	T5 _{RAED}	57.0	47.4	53.5

Table 5.4. We evaluate EL performance on temporal test sets (2020–2022) using descriptions generated by T5-Large, RAED (with T5-Large), and the gold description. Results in bold indicate the best performance for generated descriptions.

5.5.3 Emerging Entity Linking

In this section, we evaluate the effectiveness of RAED in handling emerging entities. As detailed in Section 5.4.3, we use the temporal splits from Wikipedia provided by the TempEL dataset. For each split, we define an entity as *emerging* if it did not appear in any of the previous snapshots (e.g., entities added between 2020 and 2021 for the 2021 split). Although TempEL includes splits from 2013 to 2022, we focus our analysis on those from 2019 onward to better simulate scenarios in which emerging entities are unseen during model pre-training, particularly for models such as T5.

We evaluate EL performance on temporal test sets using the TempEL model and two types of descriptions: (1) descriptions generated by RAED, and (2) descriptions generated by models trained without retrieval. The models used are the same as those reported in Table 5.1. Following Zaporozhets et al. (2022), we report accuracy@64 as our evaluation metric, which measures whether the target entity appears within the top-64 retrieved candidates. To ensure a fair evaluation and prevent data leakage, we exclude from the index any passages taken from the Wikipedia pages of the emerging entities.

TempEL (Zaporozhets et al., 2022) offers not only temporal splits of Wikipedia but also Entity Linking models trained on each yearly snapshot. For instance, TempEL provides ten bi-encoder models trained on Wikipedia snapshots from 2013 to 2022, meaning that the 2019 TempEL model was trained solely on the 2019 Wikipedia split. We use this model in

our experiments to accurately simulate the emerging entity setting and to avoid data leakage from model training. Table 5.4 shows not only the performance of the TempEL model when provided with the generated definitions in its index, but also with the gold definitions. This would be considered the ceiling for performance in Entity Description for Emerging Entity Linking. As seen in Table 5.4, RAED significantly outperforms the model with no access to retrieved knowledge, achieving an average improvement of 8.1 points. This shows that by leveraging retrieval, RAED generates descriptions with stronger semantic grounding, which in turn improves disambiguation and linking. This makes it particularly effective for the challenging task of EEL.

Since RAED improves downstream Emerging Entity Linking performance, it is important to verify that these gains are not due to patterns that would suggest data leakage, such as *verbatim reproduction* of retrieved passages, rather than the legitimate use of overlapping factual content (e.g., entity names, dates, domain-specific terms) expected in retrieval-augmented systems. To this end, we conduct an additional leakage analysis comparing the generated descriptions with the retrieved passages. First, for each generated description, we compare every generated sentence against the retrieved passages using exact string matching. This sentence-level check is designed to identify cases where the model reproduces retrieved evidence verbatim rather than synthesizing a new description from it. We find no exact sentence-level matches between the generated descriptions and the retrieved passages. Second, we manually inspect a random sample of 200 generated sentences and their corresponding retrieved passages to check whether the generated content appears to be directly copied or trivially rephrased from the evidence. This manual inspection did not reveal cases of direct copying or other patterns suggestive of data leakage. While this analysis does not rule out more subtle forms of paraphrase-level overlap, it provides evidence that the downstream gains are not driven by verbatim sentence-level leakage from the retrieved passages.

5.5.4 Qualitative Analysis

To complement our quantitative evaluation, we employ a Large Language Model (LLM) as a judge (Gu et al., 2024) to provide qualitative insights into the quality of generated descriptions. This approach leverages the LLM’s ability to assess linguistic fluency, semantic

Model	Win Rate
T5 _{RAED}	40%
Wikipedia (Gold)	60%
T5	9%
Wikipedia (Gold)	91%
T5 _{RAED}	82%
T5	18%

Table 5.5. Win rates (%) from LLM-as-a-Judge experiments comparing RAED, T5 without retrieval, and Gold Wikipedia descriptions.

coherence, and relevance without relying on explicit gold references (Bai et al., 2023). By comparing descriptions side-by-side, the LLM determines which one better represents the target entity in the given context, providing win rates for each comparison.

We used Phi-3.5-mini-instruct¹¹ (Abdin et al., 2024) as the judge model. The model is prompted in a pairwise manner to determine which of the two descriptions better aligns with a given mention, using the context of an entity from the unseen splits of the AIDA validation and test sets. These splits contain only entities that were not seen during training. An example of a prompt used for the LLM-as-a-Judge experiment is shown in Table 5.6.

Given the context:

INPUT_TEXT

which entity definition better matches the highlighted mention for entity "WIKIPEDIA_PAGE_TITLE"?

Description 1: 'Desc_1'

Description 2: 'Desc_2'

Answer 'Description 1' or 'Description 2'."

Table 5.6. Prompt used for our LLM-as-a-Judge experiments. It is important to note that the order in which the two descriptions appear is random.

For this experiment, we conduct three sets of comparisons: (1) T5_{RAED} vs. Wikipedia gold descriptions, (2) T5 without retrieval, vs. Wikipedia gold descriptions, and (3) T5_{RAED} vs. T5 without retrieval. The win rates reported in Table 5.5 reveal that, as expected, the gold descriptions perform best overall. RAED’s descriptions are strongly preferred over descriptions without retrieval with a win rate of 82%. Interestingly, RAED descriptions are even competitive with gold descriptions, achieving a 40% win rate, whereas the non-retrieval

¹¹<https://huggingface.co/microsoft/Phi-3.5-mini-instruct>

definitions win in only 9% of cases.

Finally, in Table 5.7, we present qualitative examples from our analysis of RAED, highlighting recurring error types made by models without retrieval, that RAED effectively avoids. In the first example, the T5 model hallucinates the middle name and birth date of *Julia Roberts*, whereas RAED generates factually accurate information and even includes examples of films she is known for, which are not mentioned in the gold definition. In the second example, the model trained without retrieved passages incorrectly states that *Benghazi* is the capital and largest city of Libya. In contrast, RAED, using the same underlying generator, correctly identifies it as the second-largest city and also names the actual capital, *Tripoli*. The third example illustrates a disambiguation error: T5 fails to resolve the mention correctly and mistakenly generates a definition for *Japan* instead of the intended entity, the *Japan national football team*. $T5_{RAED}$, on the other hand, correctly disambiguates the entity in question and generates an accurate definition.

5.6 Summary

Large-scale multimodal resources improve knowledge coverage and generalization for both concepts and entities, yet they remain inherently static. In real-world settings, new entities continuously emerge, existing entities evolve, and factual information changes over time, making closed-world assumptions and fixed knowledge base inventories increasingly inadequate. This chapter addressed this limitation by presenting a novel approach that leverages retrieval-augmented generation to support dynamic knowledge integration without requiring exhaustive retraining or complete entity inventories.

We introduced *RAED*, a retrieval-augmented framework for generating entity titles and definitions grounded in external evidence. Rather than relying on predefined candidate sets or static entity summaries, RAED retrieves contextually relevant passages and conditions generation directly on this real-world information, enabling the construction of informative descriptions even for entities that are absent from the knowledge base. A key design choice is a retrieval training paradigm centered on diverse, mention-level contextual evidence, which improves robustness under distribution shift and reduces dependence on restrictive sets of candidates.

Error	Factual
Input	<i>[...] The film starring Liam Neeson and Julia Roberts recounts the life [...]</i>
T5	Julia Roberts <def> Julia Anne Roberts (born November 17, 1967) is an American actress and filmmaker.
T5_{RAED}	Julia Roberts <def> Julia Fiona Roberts (born October 28, 1967) is an American actress known for her roles in films such as "Pretty Woman" and "Erin Brockovich".
Gold	Julia Roberts <def> Julia Fiona Roberts (born October 28, 1967) is an American actress.
Error	Factual
Input	<i>[...] police officers clash regularly in Benghazi At least [...]</i>
T5	Benghazi <def> Benghazi is the capital and largest city of Libya.
T5_{RAED}	Benghazi <def> Benghazi is the second-largest city in Libya, after the capital Tripoli.
Gold	Benghazi <def> Benghazi is the second-most-populous city in Libya as well as the largest city in Cyrenaica, with an estimated population of 859,000 in 2023.
Error	Disambiguation
Input	<i>Soccer Late Goals Give Japan Win Over Syria [...]</i>
T5	Japan <def> Japan is an island country in East Asia.
T5_{RAED}	Japan national football team <def> The Japan national football team represents Japan in men's international football and is controlled by the Japan Football Association, the governing body for football in Japan.
Gold	Japan national football team <def> The Japan national football team, also known by the nickname Samurai Blue, represents Japan in men's international football.
Error	Disambiguation
Input	<i>[...] although sixth seeded former champion Agassi had to wriggle out of a dangerous [...]</i>
T5	Michael Agassi <def> Michael David Agassi (born June 29, 1970) is an American former world No. 1 tennis player.
T5_{RAED}	Andre Agassi <def> Andre Kirk Agassi (born April 29, 1970) is an American former world No. 1 tennis player.
Gold	Andre Agassi <def> Andre Kirk Agassi (born April 29, 1970) is an American former world No. 1 tennis player.

Table 5.7. Qualitative examples comparing descriptions generated by RAED and the model without retrieval. For each example, we indicate the type of error, highlighting incorrect parts in red and corrected sections in green.

Empirical results show that *RAED* models consistently outperform non-retrieval baselines in both in-domain and out-of-domain settings, including scenarios explicitly targeting

emerging entities. In particular, evaluations on TempEL demonstrate that RAED effectively bridges the gap between static knowledge bases and evolving entity representations, while qualitative analysis confirms that retrieved evidence supports the generation of informative and factually grounded descriptions. These findings indicate that RAED can be effectively employed to automatically create and update entity entries as new information becomes available, enabling knowledge bases to remain current and reliable for downstream use.

At the same time, automatic generation of entity descriptions introduces a critical dependency on reliable validation. Even when grounded in retrieved knowledge, generated content may contain unsupported or partially incorrect statements that must be identified before integration into downstream systems. This observation motivates the final chapter of the dissertation, which focuses on interpretable and document-grounded factuality evaluation methods for validating automatically generated knowledge.

Chapter 6

Toward Interpretable Knowledge Evaluation

6.1 Overview

The previous chapters addressed two complementary challenges in knowledge-intensive NLP: expanding the coverage of multimodal knowledge representations and enabling models to incorporate emerging information through retrieval-augmented generation, thereby extending previously static knowledge inventories over time. While these advances improve a system’s ability to access, generate, and integrate knowledge in dynamic environments where new entities continuously appear, they remain valuable only if the resulting content is factually correct, evidence-grounded, and suitable for integration into downstream systems.

In particular, when automatically generated entries are incorporated into a knowledge base, even minor inaccuracies can propagate through downstream pipelines. Reliable validation is therefore a fundamental requirement for any approach that aims to maintain or expand structured knowledge at scale. Standard evaluation practices for Natural Language Generation (NLG) predominantly rely on surface-level similarity to a gold reference. While effective for measuring lexical overlap, these metrics are poorly aligned with factual correctness, especially in open-ended and knowledge-intensive settings. For example, in the context of emerging or evolving entities, reference descriptions are often unavailable, incomplete, or inherently subjective. Even when a reference exists, it represents only one possible summary of a much larger evidence source. As a result, a generated description

may introduce correct and useful information that is absent from the reference, or conversely, achieve high lexical similarity while containing unsupported or incorrect claims.

This limitation is especially evident for Wikipedia-derived entity descriptions, which function as summaries of entire articles rather than fixed reference texts. Different authors may emphasize different aspects of the same entity, and valid descriptions may vary substantially in both content and structure. Evaluating such outputs therefore requires assessing whether the generated statements are supported by the underlying source evidence, rather than whether they resemble a specific target string. This perspective aligns closely with factuality evaluation in abstractive summarization, where the central challenge is to verify that generated content is supported by the source document.

With this motivation, this chapter introduces *FENICE*, an evaluation framework designed for reference-free factuality assessment by validating model outputs against source evidence at the level of atomic claims. We then show how *FENICE* serves as a practical and more complete validation layer for *RAED*: it enables automatic verification of newly generated entity descriptions before they are used to extend existing inventories, and it highlights precisely which statements are unsupported and why.

In summary, the contributions of this chapter are as follows:

- We introduce *FENICE*, an interpretable evaluation framework that enables *reference-free* factuality assessment by decomposing generated text into atomic claims and verifying each claim against source documents using natural language inference.
- We apply *FENICE* to *RAED* outputs to validate emerging knowledge before inventory updates, providing fine-grained analyses that identify which generated claims are supported, unsupported, or contradicted by the retrieved evidence.
- We design *FENICE* to return both quantitative scores and explicit evidence attribution, enabling downstream knowledge inventories to incorporate only factually validated entries and supporting transparent auditing of newly introduced entity information.

We also find that *FENICE* generalizes beyond emerging-knowledge validation, yielding strong performance on factuality evaluation for summarization. By enabling evidence-grounded and interpretable validation of model-generated knowledge, this chapter addresses

a key validation requirement for reliable and scalable knowledge-intensive systems operating under continuously evolving real-world information.

6.2 Factuality Evaluation in Knowledge-Intensive Generation

Recent advances in NLG have led to substantial improvements across a wide range of NLP tasks, including machine translation, question answering, and text summarization. In many of these settings, models are now capable of producing fluent, coherent, and contextually appropriate outputs. However, fluency alone is insufficient in knowledge-intensive applications, where correctness depends on faithful grounding in external information. This limitation is particularly evident in text summarization, where empirical studies have consistently shown that a non-negligible portion of generated summaries contains factual inconsistencies or unsupported statements, despite achieving high scores under surface-level evaluation metrics such as ROUGE (Lin, 2004). Multiple empirical investigations have confirmed this phenomenon, revealing that a significant fraction of automatically generated summaries exhibits factual errors and hallucinations (Maynez et al., 2020; Durmus et al., 2020; Kryscinski et al., 2020).

The emergence of Large Language Models (LLMs) has yet to resolve this issue. While LLMs substantially improve linguistic quality, their outputs continue to exhibit hallucinations and factual errors (Tam et al., 2023). More importantly, their increased fluency and confidence can obscure these errors, encouraging users to over-trust generated content and facilitating the propagation of misinformation (Wang et al., 2023b). As a result, reliable factuality evaluation has become a prerequisite for deploying generation systems in real-world, knowledge-sensitive settings.

Automatic Factuality Metrics. In response to these challenges, a growing body of work has proposed automatic metrics for factuality evaluation. Early approaches draw on Question Answering paradigms, assessing factual consistency by checking whether a summary enables correct answers to questions derived from the source document (Wang et al., 2020; Scialom et al., 2021; Fabbri et al., 2022). Other methods leverage Natural Language Inference (NLI) to verify whether the generated text is entailed by the input document (Goyal and Durrett, 2020; Laban et al., 2022; Zha et al., 2023; Chen and Eger, 2023). More recent approaches

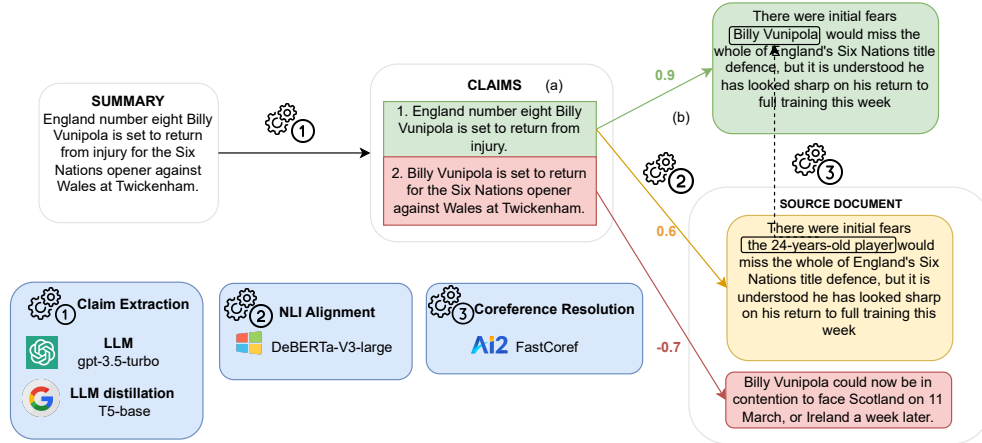


Figure 6.1. Overview of FENICE: the process begins with the extraction of claims from a given summary (step 1). Extracted claims are then aligned with specific sections of the input document (step 2). Finally, we refine the obtained alignments through a coreference-resolution-based approach (step 3). Best seen in color.

incorporate semantic graph representations or structured fact comparisons to capture deeper relationships between the source and the summary (Ribeiro et al., 2022; Fan et al., 2023).

Despite steady progress, existing metrics share several persistent limitations. First, most methods assign a single score to the entire output, offering limited insight into which specific statements are factual or incorrect. Second, many metrics are tailored to short documents, such as news articles, and struggle to scale to longer or more complex inputs. Finally, a growing class of LLM-based metrics achieves strong correlation with human judgments but is computationally expensive, difficult to reproduce, and often dependent on closed-source models (Fu et al., 2024; Liu et al., 2023b; Gao et al., 2023).

Interpretability and Evidence Attribution. A central shortcoming of many existing metrics is their lack of interpretability. NLI-based methods such as MENLI (Chen and Eger, 2023) and AlignScore (Zha et al., 2023) evaluate factuality by comparing the entire summary against the source document, but provide limited diagnostic feedback. When inconsistencies are detected, users are not told which parts of the generated text are unsupported, nor which portions of the source document are relevant for verification. Similarly, approaches based on dependency structures or sentence-level entailment struggle to handle the highly abstractive nature of modern summaries, where multiple facts may be fused into a single sentence (Goyal and Durrett, 2020).

This limitation becomes especially problematic when generated content is intended for downstream integration, such as knowledge base population or entity description generation. In such settings, factuality evaluation must not only detect errors, but also localize them and provide explicit evidence that supports or refutes individual claims.

LLMs as Evaluators and Their Limitations. Recent work has explored the use of LLMs as factuality judges, either directly through prompting or indirectly through data generation and distillation (Shen et al., 2023; Gekhman et al., 2023b; Fu et al., 2023; Liu et al., 2023b). While these approaches often outperform traditional metrics in correlation with human judgments, they introduce new concerns. Their reliance on closed-source models raises uncertainty about data contamination, and their computational cost limits applicability in large-scale or real-time evaluation scenarios. Moreover, using the same class of models for both generation and evaluation risks masking systematic errors rather than exposing them.

Claim-Level Evaluation. An emerging direction tackles the evaluation of factuality at the level of atomic claims. Instead of scoring entire outputs, these approaches decompose generated text into individual factual statements and verify each one against an external source (Min et al., 2023). This formulation naturally supports interpretability, enables fine-grained diagnostics, and aligns more closely with how humans assess factual correctness. However, existing claim-based methods are often designed for specific tasks, rely on external knowledge bases rather than source documents, or lack efficient and scalable implementations.

These observations motivate the need for an evaluation framework that combines the strengths of claim-level analysis with document-grounded verification, computational efficiency, and explicit evidence attribution. In the remainder of this chapter, we introduce FENICE, a factuality evaluation framework designed to meet these requirements and to support reliable validation of generated knowledge, including in dynamic settings involving emerging entities and retrieval-augmented generation.

6.3 FENICE

We now provide an overview of the FENICE algorithm. Initially, we extract a list of atomic facts, referred to as claims, from a given summary (Section 6.3.1). Subsequently, we detail the process of obtaining a factuality score for each extracted claim. This involves leveraging NLI, with the claim serving as the hypothesis, and the document sentences (Section 6.3.2) or longer blocks of text at different levels of granularity (Section 6.3.4) serving as premises. In Section 6.3.3, we elaborate on a refinement of the obtained scores through coreference resolution.

6.3.1 Claim extraction

Following the Atomic Content Unit (ACU) definition outlined in Liu et al. (2023d), we use the term *claim* to denote an *atomic fact*, i.e., an elementary information unit found in a summary, that does not require further subdivision. Figure 6.1(a) depicts an example list of claims extracted from a summary.

We frame the claim extraction problem as an end-to-end autoregressive generation task. Given a prompt consisting of the task description and the summary to be processed, the model generates a sequence of claims expressed within the provided summary. To guide the model in meeting the desired output format and enhancing its task comprehension, we include an example of an input summary and the expected claims in the prompt. Figure 6.2 illustrates the formulated prompt designed for extracting claims from a given summary. This prompt was employed to generate the claims utilized in our evaluation metric and to train our distilled claim generation model.

We opt to center our evaluation strategy around claims rather than sentences, primarily because of their enhanced interpretability and their atomic nature, which facilitates alignment with information from the input.

6.3.2 NLI-based claim scoring

The NLI formulation lies at the core of our novel claim scoring algorithm. NLI is a task that involves determining the logical relationship between two text segments: a *premise* and a

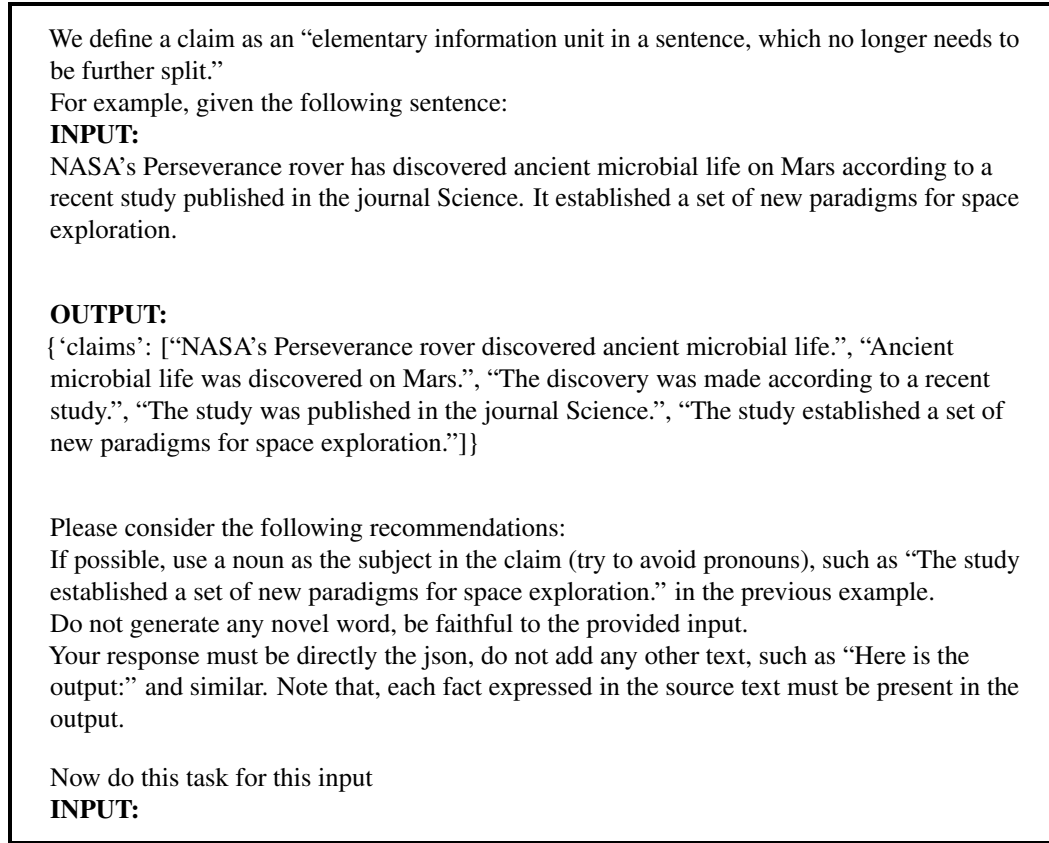


Figure 6.2. LLM prompt for extracting claims given a summary.

hypothesis. Formally, given a premise *pre* and a hypothesis *hyp*:

$$\text{NLI}(pre, hyp) = (p_{ent}, p_{neu}, p_{con}) \quad (6.1)$$

where p_{ent} , p_{neu} , and p_{con} denote the respective probabilities of *pre* entailing, being neutral about, or contradicting *hyp*.

Starting from Equation 6.1, we define NLIScore as the difference between the entailment and contradiction probabilities, obtained using the document sentence as premise and the claim extracted from the summary as hypothesis:

$$\text{NLIScore}(s, c) = p_{ent} - p_{con}$$

where the premise *s* is a sentence of the input document, and the hypothesis *c* is a claim extracted from the summary. NLIScore measures the degree to which a specific premise can

be utilized to verify the claim c . The inclusion of the contradiction probability serves the purpose of penalizing claims that are in conflict with any content in the source document.

Finally, we introduce FENICE_{sent} as the function that assigns a score to a given claim c by considering the sentences in the source document D as premises:

$$\text{FENICE}_{sent}(D, c) = \max_{s \in D} \text{NLIScore}(s, c) \quad (6.2)$$

We also define align_{sent} as the corresponding alignment function:

$$\text{align}_{sent}(D, c) = \underset{s \in D}{\text{argmax}} \text{NLIScore}(s, c) \quad (6.3)$$

For each claim c from the summary, this establishes a one-to-one alignment between c and the highest-scoring sentence from the input document.

6.3.3 Coreference Resolution

Importantly, given a claim c and its aligned sentence $s = \text{align}_{sent}(D, c)$, $\text{NLIScore}(s, c)$ may not reflect the true degree of entailment for the (s, c) pair due to an entity or concept being mentioned in the two texts with distinct surface forms. For instance, consider the claim in Figure 6.1(a): "*England number eight Billy Vunipola is set to return from injury*" and its aligned source document sentence: "*There were initial fears the 24-years-old player would miss [...]*". In this case, the NLI model assigns only a moderate score of 0.6 to this pair due to the limited context, making it challenging to establish a clear correspondence between "Billy Vunipola" and "the 24-years-old player".

To tackle this challenge, we apply a Coreference Resolution model on the source document, which identifies clusters of mentions that refer to the same entity or concept. We then introduce a coreference-resolution-aware version of Equation 6.2 as follows:

$$\begin{aligned} s^* &= \text{align}_{sent}(D, c) \\ \text{FENICE}_{coref}(D, c) &= \max_{s' \in \text{Co}(s^*)} \text{NLIScore}(s', c) \end{aligned} \quad (6.4)$$

where $\text{Co}(s^*)$ is the set of sentences obtained by replacing each entity occurring in sentence

s^* with all its coreferential mentions.¹

For instance, given "the 24-years-old player" we identify its coreferential spans in the document, such as "Billy Vunipola". Subsequently, we generate an alternative version of the original sentence by replacing "the 24-years-old player" with "Billy Vunipola", resulting in the sentence: *"There were initial fears Billy Vunipola would miss [...]"*. As depicted in Figure 6.1(b), this approach facilitates the alignment process of the claim *"England number eight Billy Vunipola is set to return from injury"* to the revisited version of the sentence, which contains the same span for the given named entity.

6.3.4 Aligning claims across multiple input text granularities

Given the abstract nature of summaries, the verification of a claim may require a context that extends beyond a single sentence from the input. For example, to verify the claim: *"Regular exercise contributes to overall well-being"*, the following input passage would be necessary: *"Studies have consistently shown that individuals who engage in regular physical activity experience improved mental health. Furthermore, regular exercise is linked to a reduced risk of chronic diseases and enhanced cardiovascular health."*

Therefore, we define the following variant of Equation 6.2 for longer passages of k sentences:

$$\text{FENICE}_p(D, c, k) = \max_{i=1, \dots, |D|-k+1} \text{NLIScore}(s_i^{i+k-1}, c) \quad (6.5)$$

where s_i^{i+k-1} is the concatenation of $s_i, \dots, s_{i+k-1}$, i.e., the k consecutive sentences starting from sentence i in document D .

Our approach involves choosing k such that $k < |D|$, alongside considering the entire document as premise, i.e., $k = |D|$. Thus, our function for scoring claims given passages of multiple lengths from the source document is given by the following:

$$\text{FENICE}_{mul}(D, c, j) = \max\{\text{FENICE}_p(D, c, j), \text{FENICE}_p(D, c, |D|)\} \quad (6.6)$$

where j is the desired length for the contiguous blocks of the source sentences. Then, we

¹We limit this augmentation to the sentence s^* , which is the result of Equation 6.3.

define FENICE_{cl} as the factuality score assigned to a claim c :

$$\text{FENICE}_{cl}(D, c, j, T) = \begin{cases} \text{FENICE}_{mul}(D, c, j) \\ \text{if } \text{FENICE}_{coref}(D, c) < T; \\ \text{FENICE}_{coref}(D, c) & \text{otherwise.} \end{cases} \quad (6.7)$$

where T is a threshold. We introduce T to limit the application of FENICE_{mul} to only a subset of claims that requires further verification, i.e., the low-scoring ones. We set $j = 5$ and $T = 0.8$ in all our experiments (Section 6.4.2) as a result of tuning them on the validation splits.

Finally, given the set of extracted claims C from a summary S , and a source document D , we introduce FENICE as:

$$\text{FENICE}(D, C, j, T) = \frac{\sum_{c \in C} \text{FENICE}_{cl}(D, c, j, T)}{|C|} \quad (6.8)$$

which is the factual consistency score attributed by our metric to summary S .

6.4 Experiments and Results

6.4.1 Claim extraction

Experimental setup. As detailed in Section 6.3.1, we formulate the claim extraction problem as a sequence generation task. To accomplish this, we designed a prompt that includes the task description, an input-output example, and the summary to be processed. We experiment with *GPT-3.5-turbo* as claim extractor.² We additionally provide a distilled version of the claim extractor, which consists of a T5_{base} (Raffel et al., 2019) model trained on the LLM outputs. To create the training and evaluation datasets for the distilled model, we apply the LLM-based claim extractor on the reference summaries from two widely-used summarization datasets, namely XSum (Narayan et al., 2018) and CNN/DailyMail (Nallapati et al., 2016). We split the dataset obtained into training, validation, and test sets following a 80%-10%-10% proportion, comprising 430,946 summary-claims pairs for training, and 53,868 for both the validation and test partitions. We train our system on a single GPU

²We use the *gpt-3.5-turbo* API, version dated 2nd December 2023.

NVIDIA GeForce RTX 3090 for a total of 1M steps, using AdaFactor (Shazeer and Stern, 2018b) as optimizer with learning rate of $1 \cdot 10^{-5}$.

Datasets. We evaluate our distilled version of the claim extractor on the test split of the LLM-generated claim extraction dataset. We additionally assess the performance of both the LLM-based claim extractor and our distilled version on ROSE (Liu et al., 2023c), a dataset that includes summaries and human-written claims among its annotations. While this dataset was designed to serve as the basis for a human annotation protocol for summarization evaluation, we utilize the released manually-extracted claims for assessing our systems’ performance.

Metrics. Given the extractive nature of this task, we consider ROUGE (Lin, 2004) as the basis for a suitable evaluation metric for claim extraction. Following (Zhang and Bansal, 2021) we define easiness_P as the score obtained by computing the ROUGE-1 metric for each pair of system-generated and gold standard claims, subsequently calculating the maximum, and averaging over the total number of generated claims:

$$\text{easiness}_P(S, H) = \frac{\sum_{c \in S} \max_{c^* \in H} R1(c, c^*)}{|S|} \quad (6.9)$$

where $R1$ is the ROUGE-1-F1 metric, and S, H are the sets of system- and human-written claims, respectively. We then introduce easiness_R as the recall-oriented version of Equation 6.9, obtained by averaging over the gold standard claims:

$$\text{easiness}_R(S, H) = \frac{\sum_{c^* \in H} \max_{c \in S} R1(c, c^*)}{|H|} \quad (6.10)$$

Results and discussion. In Table 6.1 we report the results of the easiness_P , and easiness_R metrics, along with the F1 version of Equations 6.9 and 6.10, i.e., easiness_{F1} , obtained on ROSE. Additionally, in Table 6.3, we report the claim extraction evaluation of the LLM-based claim extractor compared to its T5-based distilled version, using BERTScore instead of ROUGE.

As we can see from Table 6.1, our distilled version performs comparably with the original one based on *GPT-3.5*. These results indicate the effectiveness of the distillation

Model	easiness _P	easiness _R	easiness _{F1}
GPT-3.5	80.1	70.9	74.9
T5 _{dist_GPT}	79.2	68.8	73.4

Table 6.1. Easiness Precision (easiness_P), Recall (easiness_R), and F1 score (easiness_{F1}) results for the LLM-based claim extractor, namely *GPT-3.5*, and its distilled version *T5_{dist_GPT}*, assessed on ROSE (Liu et al., 2023d).

Model	easiness _P	easiness _R	easiness _{F1}
T5 _{dist_GPT}	89.7	89.9	89.5

Table 6.2. Easiness metric results of the distilled version of the LLM-based claim extractor, i.e., *T5_{dist_GPT}* evaluated on the test set of the dataset composed of claims generated by *GPT-3.5*.

process, ensuring that our model can be used reliably for extracting claims. To further evaluate the quality of the claims generated by our distilled model, we compare them to those generated by the original GPT-3.5 model, using the latter as references. The results of this evaluation are presented in Table 6.2, demonstrating the strong ability of our T5 model to replicate the claims generated by GPT-3.5.

6.4.2 Factuality Evaluation

Experimental setting. We leverage an off-the-shelf model trained on NLI to align the summary-extracted claims with text from the input document. Specifically, we employ a version of DeBERTa-v3-large (He et al., 2020) fine-tuned on multiple NLI datasets,³ comprising MultiNLI (Williams et al., 2017), LingNLI (Parrish et al., 2021), WANLI (Liu et al., 2022), and Fever-NLI, as well as Adversarial-NLI, both introduced in Nie et al. (2019). We adopt spaCy’s *en-core-web-sm* to segment the input documents into sentences, and F-Coref (Otmazgin et al., 2022) for the extraction of the coreferential spans (Section 6.3.3).

Evaluation datasets. We assess the effectiveness of FENICE on AGGREGFACT (Tang et al., 2023), the de-facto benchmark for evaluating summarization factuality metrics. AGGREGFACT comprises multiple datasets of news articles and manually annotated summaries with factuality labels. The benchmark is categorized into three splits based on the summarization systems employed: i) FTSOTA, which includes the outputs of state-of-the-art pretrained

³We loaded the pre-trained model weights from the Hugging Face Hub, available at: <https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli>

Model	Precision (%)	Recall (%)	F1 (%)
GPT-3.5	95.8	94.3	95.0
T5 _{dist_GPT}	95.5	93.9	94.7

Table 6.3. Claim extractor performance comparison on ROSE (Liu et al., 2023d) using BERTScore (Zhang et al., 2020) as backbone metric in Equations 6.9 and 6.10 instead of ROUGE.

summarizers, ii) EXFORMER, consisting of summaries of early Transformer-based summarization models, and iii) OLD, which comprises the outputs of older approaches.

Baselines. We evaluate against several NLI-based metrics, including AlignScore (Zha et al., 2023), MENLI (Chen and Eger, 2023), SummaC-ZS and SummaC-Conv (Laban et al., 2022), and DAE (Goyal and Durrett, 2020). For QA-based evaluation, we select QAFactEval (Fabbri et al., 2022), the top-performing QA metric on AGGREFACT, which relies on a Question Generation model to formulate questions from the source document and checks whether the summary can answer them, and vice versa. Additionally, we benchmark several LLM-based metrics: ChatGPT-ZS and ChatGPT-CoT (Luo et al., 2023b), which leverage GPT-3.5 to generate a binary factuality label in zero-shot and Chain-of-Thought paradigms, respectively, and ChatGPT-DA and ChatGPT-Star (Wang et al., 2023a), which prompt LLMs to generate factuality scores on a scale of 0–100 and 1–5, respectively. We also include TrueTeacher (Gekhman et al., 2023b), specifically its 11B-parameter T5 variant.

Results and discussion. We assess the performance of two distinct variants of our metric: FENICE_{GPT_claims}, which employs LLM-based claims, and FENICE_{T5_claims}, featuring the knowledge-distilled claim extractor. To convert FENICE scores into binary labels, we tune a different threshold on each validation split of the AGGREFACT datasets (i.e., FTSOTA, EXFORMER, and OLD) and then compute the balanced accuracy on the corresponding test set, as indicated in Tang et al. (2023).

In Table 6.4, we present the balanced accuracy (Brodersen et al., 2010) scores on the AGGREFACT test sets. FENICE_{GPT_claims} sets a new state-of-the-art performance on AGGREFACT, resulting in the highest average balanced accuracy score.

Moreover, we specifically analyze the metric’s ability to evaluate the outputs of state-of-the-art models. Following the recommendation of Tang et al. (2023), we report

	Agg-CNN			Agg-XSum			AVG
	FTS	EXF	OLD	FTS	EXF	OLD	
Random Baseline	50.0	50.0	50.0	50.0	50.0	50.0	50.0
DAE*	59.4	67.9	69.7	73.1	-	-	67.5
QuestEval	63.7	64.3	65.2	61.6	60.1	59.7	63.5
SummaC-ZS	63.3	76.5	76.3	56.1	51.4	53.3	65.3
SummaC-Cv	70.3	69.8	78.9	67.0	64.6	67.5	70.5
TrueTeacher-11B	65.7	57.7	<u>81.9</u>	<u>75.2</u>	68.4	52.8	67.0
QAFactEval	61.6	69.1	80.3	65.9	59.6	60.5	67.0
MENLI	51.7	52.8	68.4	58.3	59.7	73.9	60.8
AlignScore	53.5	73.9	78.0	80.2	79.9	63.7	<u>71.5</u>
ChatGPT-ZS	66.2	64.5	74.3	62.6	69.2	60.1	65.7
ChatGPT-CoT	49.7	60.4	66.7	56.0	60.9	50.1	57.9
ChatGPT-DA	48.0	63.6	71.0	53.6	65.6	61.5	60.1
ChatGPT-Star	55.8	65.8	71.2	57.7	70.6	53.8	63.8
FENICE _{GPT_claims}	<u>68.2</u>	68.8	82.1	73.9	<u>73.5</u>	<u>69.9</u>	72.7
FENICE _{T5_claims}	61.6	67.8	80.6	71.4	70.7	67.7	70.0

Table 6.4. Balanced accuracy results on the test sets of the AGGREFACT datasets. Following Tang et al. (2023) we exclude the performance of DAE on the EXFORMER (EXF in table) and OLD datasets in the AGGREFACT-XSUM partition, since it was trained on XSumFaith (Goyal and Durrett, 2021) which is part of those splits. Results in **bold** indicate the best performance, while underlined values indicate the second best.

the results of our metrics on AGGREFACT-FTSOTA split in the single-threshold setting. As outlined by Tang et al. (2023), we fine-tune a threshold on the validation sets of the CNN-FTSOTA and XSUM-FTSOTA splits and subsequently apply it to convert our metric scores into binary labels on the test set. From the outcomes presented in Table 6.5, we note that FENICE_{GPT_claims} obtains the best average results in this split, also due to its high performance on XSUM-FTSOTA. Notably, we consider this latter setting to be the most challenging for a metric, given that it comprises summaries of state-of-the-art systems trained on XSum references, which exhibit the highest level of abstraction. In this context, FENICE_{T5_claims} slightly outperforms TrueTeacher, despite utilizing significantly fewer parameters (700M vs 11B).

6.4.3 Ablation study

We conduct an analysis of the impact of our primary design choices, namely, claim extraction, coreference resolution, the alignment with multiple text granularities, and our backbone NLI

Metric	CNN(FTS)	XSum(FTS)	AVG
DAE	65.4±4.4	70.2±2.3	67.8
QuestEval	70.2±3.2	59.5±2.7	64.9
SummaC-ZS	64.0±3.8	56.4±1.2	60.2
SummaC-Conv	61.0±3.9	65.0±2.2	63.0
QAFactEval	67.8 ±4.1	63.9±2.4	65.9
TrueTeacher-11B	62.0±1.3	74.9±1.2	68.4
MENLI	63.4±2.8	59.0±1.8	61.2
AlignScore	62.7±3.1	69.4±1.9	66.1
ChatGPT-ZS	56.3±2.9	62.7±1.7	59.5
ChatGPT-COT	52.5±3.3	55.9±2.1	54.2
ChatGPT-DA	53.7±3.5	54.9±1.9	54.3
ChatGPT-Star	56.3±3.1	57.8±0.2	57.1
FENICE _{GPT_claims}	70.5±1.8	<u>72.8±0.3</u>	71.6
FENICE _{T5_claims}	67.7±3.0	70.1±1.8	<u>68.9</u>

Table 6.5. Balanced accuracy results on the AGGREFACT-FTSOTA split obtained with the single-threshold setting. Results in **bold** indicate the best performance, while underlined values represent the second best.

model.

First, we introduce NLI_{sent} , a simple baseline leveraging NLI-based sentence-level alignment between the summary and the input document. Specifically, the factuality score for a summary sentence s' is obtained through the following variant of Equation 6.2, considering summary sentences as hypotheses instead of claims:

$$NLI_{sent}(D, s') = \max_{s \in D} NLI_{Score}(s, s') \quad (6.11)$$

As outlined in Section 6.3.4, we obtain the factuality score for a summary by averaging the result of Equation 6.11 applied to its sentences.

We then assess the impact of focusing the factuality evaluation on claims extracted from the summary rather than its sentences. To achieve this, we introduce NLI_{claim} , a modification of the previous baseline that assigns factuality scores to the claims using Equation 6.2. Subsequently, we evaluate the performance of NLI_{coref} , derived by integrating the coreference resolution strategy (Section 6.3.3) into the previous baseline. Additionally, we assess the impact of including the alignment strategy with input text at multiple granularities (Section 6.3.4) by presenting the performance of FENICE. Table 6.6 displays the balanced

Metric	CNN	XSUM	AVG	Δ AVG
NLI_{sent}	52.2 $\pm$ 1.4	59.6 $\pm$ 0.6	55.9	–
NLI_{claim}	54.3 $\pm$ 2.1	63.2 $\pm$ 0.8	58.8	+2.9
NLI_{coref}	59.7 $\pm$ 2.3	63.7 $\pm$ 0.9	61.5	+2.7
FENICE	70.5$\pm$3.0	72.8$\pm$0.3	71.6	+10.1

Table 6.6. Ablation study evaluating the impact of component additions on FENICE on the AGGREGFACT-FTSOTA split. We report the results of the following baselines: i) NLI_{sent} , utilizing summary sentences instead of claims, ii) NLI_{claim} , which leverages claims, and iii) NLI_{coref} , obtained by integrating the coreference-resolution-based alignments.

Metric	CNN	XSUM	AVG	Δ AVG
FENICE	67.5$\pm$3.0	73.7$\pm$0.3	70.6	–
FENICE _{RoBERTa}	64.4 $\pm$ 3.2	68.1 $\pm$ 1.6	66.3	-4.3

Table 6.7. Ablation study. We compare the performance of FENICE on the AGGREGFACT-FTSOTA split using DeBERTa and RoBERTa as the backbone NLI model.

accuracy scores of FENICE on the AGGREGFACT- FTSOTA split, comparing them to the aforementioned baselines.

The values in the ΔAVG column represent the differences in terms of average balanced accuracy resulting from the incremental inclusion of each component. The results indicate that each component contributes significantly to the final performance. The alignment strategy with input text at multiple granularities has the most substantial impact, resulting in a significant gain of 10.1 points when included. Additionally, we evaluate the performance of FENICE utilizing RoBERTa⁴ (Liu et al., 2019) as the NLI model for our alignments (FENICE_{RoBERTa}), instead of DeBERTa, as shown in Table 6.7.

6.4.4 Long-form summarization evaluation

To assess whether our proposed metric remains effective in realistic knowledge-intensive settings, we extend the evaluation to long-form documents. This scenario is particularly relevant to the main goal of this chapter, which is evaluating emerging entity knowledge. In practice, newly emerging entities often lack dedicated curated pages, meaning that evaluation must rely on large, heterogeneous documents retrieved as evidence rather than focused, entity-specific documents. As a result, factual verification must operate over longer contexts

⁴We load the pre-trained model weights from the Hugging Face Hub, available at: `ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli`.

that contain dispersed and partially relevant information.

This requirement aligns with a broader shift in summarization research toward long-form content. Recent work has increasingly focused on summarizing extended documents such as scientific articles (Gupta et al., 2021) and books (Kryściński et al., 2022; Scirè et al., 2023), reflecting the need to process richer and more complex information sources. However, evaluation methodologies have evolved more slowly, with only a limited number of efforts explicitly addressing how factual consistency should be assessed in long-document settings (Krishna et al., 2023).

Within the context of this thesis, long-form evaluation is not merely a summarization challenge but a necessary condition for validating automatically generated entity descriptions grounded in retrieved evidence. By demonstrating that our metric can reliably assess factual consistency when evidence spans large documents, we show that it supports realistic evaluation scenarios in which knowledge is incomplete, evolving, and distributed across multiple sources.

Human evaluation. As an initial step towards addressing this gap, we conducted an evaluation of our metric in the context of long-form text summarization. To do so, we curated a set of annotations for assessing summaries of lengthy texts.

Drawing upon Echoes from Alexandria (Scirè et al., 2023), a multilingual resource for book summarization, we selected 26 English narratives, each paired with a reference abstractive summary. The reference summaries were refined by removing irrelevant details such as the story title, the genre, and the author’s name. As an example, the initial summary *"Isaac Asimov’s ‘Let’s Get Together’ is a science fiction novel which depicts the continuation of Cold War hostilities..."* underwent manual editing and was refined to *"The story depicts the continuation of Cold War hostilities..."*. To augment our selection, we incorporated summaries generated by 12 book summarization systems, as presented by Scirè et al. (2023), applied to our selected stories. The systems chosen were versions of BART (Lewis et al., 2020a), LongT5 (Guo et al., 2022), and LED (Beltagy et al., 2020), fine-tuned on the *Echo-XSum* split of the *Echoes from Alexandria* dataset. This process resulted in a comprehensive set of 338 text-summary pairs.

Next, we tasked an expert annotator with reading all the stories and assigning a binary

Dataset	Avg. Sentences	Avg. Tokens
Ours	664.2	11,817.9
SummEval	18.0	363.1
SummaC	23.0	485.5

Table 6.8. Dataset Statistics.

Factual	Not Factual
66	273

Table 6.9. Number of factual/unfactual summaries in our set of human annotations.

factuality label to each of the associated summaries.⁵ As the foundation for the annotation process, we considered a summary factual if its content was explicitly stated or implied by the source document; otherwise, it was labeled as non factual. Through this process, we acquired factuality labels for all the 338 text-summary pairs.

Moreover, in order to measure the inter-annotator agreement we asked a second expert to manually annotate a subset of 52 summaries. We then computed the Cohen’s kappa coefficient, obtaining a moderate agreement of 0.74. We deemed this outcome reasonable, particularly when considering the high complexity of assessing long-form summarization compared to the standard setting.

Table 6.8 compares the length of source texts in our proposed long-form summarization evaluation dataset with those in existing datasets, specifically SummaC (Laban et al., 2022) and SummEval (Fabbri et al., 2021). In our dataset, annotators were tasked with reading very long source texts, averaging 664 sentences, in order to assess the factuality of the provided summaries. Of the 339 summaries annotated, 66 (19.5%) were judged factual and 273 (80.5%) were judged non-factual. In Table 6.9, we report the number of factual and unfactual summaries in our collection.

To align with the evaluation setup of AGGREFACT, we partitioned our annotations into test and validation sets using an 80/20 split.

⁵We supplied the annotator with the PDF files containing the text of the stories. We paid the annotator according to the standard salary for their geographical location. The annotator took an average of 2 hours to read a story, and an average of 5 minutes to annotate the corresponding summaries.

Metric	Balanced Accuracy
AlignScore	61.3
MENLI	61.7
DAE	51.4
FENICE _{GPT_claims}	66.2
FENICE _{T5_claims}	<u>65.7</u>

Table 6.10. Balanced accuracy results of FENICE metrics and the selected baselines on our set of human annotations for long form summarization factuality. Results in **bold** indicate the best performance, while underlined values represent the second best.

Results and discussion. We follow the experimental setup outlined in Section 6.4.2. We employ AlignScore, MENLI, and DAE as baselines, considering their relatedness to our work. In Table 6.10, we present the balanced accuracy results on the test split of our dataset. The results obtained highlight that FENICE metrics outperform the baselines, with FENICE_{GPT_claims} obtaining the highest performance and FENICE_{T5_claims} coming second. Upon manual inspection of our alignments, we observe that a significant proportion of claims are aligned with source text having the length of a paragraph. We posit that this behaviour is due to the high abstractiveness of the stories’ summaries, which typically require contexts that are longer than a sentence in order to be verified. We hypothesize that our metric is well-suited for both standard and long-form summarization thanks to its adaptive capacity to choose the appropriate input text granularity for claim verification.

6.5 FENICE for Interpretable Entity Description Evaluation

While factuality metrics such as F-NLI provide a quantitative score of whether a generated text is supported by a reference description, they do not reveal which specific statements are unsupported or incorrect. Furthermore, F-NLI can only operate in the presence of a gold reference description, which is typically unavailable for emerging entities in real-world settings. FENICE addresses these limitations by providing (i) claim-level verification through atomic fact decomposition and (ii) document-grounded verification that does not rely on a gold reference description. Concretely, FENICE extracts atomic claims from the generated text and verifies each claim individually against the source document using natural language inference, producing fine-grained factuality labels together with explicit evidence attribution.

In our setting, Wikipedia descriptions can be viewed as summaries of the corresponding entity pages (i.e., the lead section summarizes the most important contents of the article). We therefore reformulate factuality evaluation of generated entity descriptions as a summarization factuality problem. In order to do so, we retrieve the plain text of the Wikipedia article for each target entity in our test sets and remove the entity description (lead) section to prevent contamination/leakage, ensuring that the *document* does not trivially include the same content as the *summary* being verified. As an additional leakage control, we further check the retained article text for direct textual overlap with the removed entity description. Specifically, we verify that the retained document does not contain exact sentence-level matches with the removed lead section, nor exact contiguous spans longer than five words. This check is intended to distinguish legitimate evidential support from direct textual reuse: the retained article may still contain factual details needed for verification, such as entity names, dates, occupations, locations, or domain-specific terminology, but it should not simply reproduce the removed description. While this procedure does not rule out all possible paraphrase-level overlap, it reduces the risk that FENICE validates generated descriptions against evidence that directly restates the reference description. We treat the remaining article text as the *document* input to FENICE, and the RAED-generated description as the *summary* to be verified. For this reason, we apply FENICE to the outputs of RAED models to assess whether retrieval-augmented generation improves factual grounding beyond what is captured by reference-based and sentence-level metrics. We also compare the factuality scores produced by FENICE to those of F-NLI, which has access to gold reference descriptions, to assess whether reference-free, document-grounded evaluation yields consistent signals.

Table 6.11 reports the comparison between F-NLI and FENICE across in-domain and out-of-domain evaluation settings. Both metrics exhibit consistent patterns: models that benefit from retrieval achieve higher factuality scores, and these gains persist in out-of-domain settings where baseline performance degrades. This agreement is notable because FENICE operates without access to any gold reference description, supporting its applicability in settings where gold references are unavailable, such as for emerging or rapidly evolving entities. FENICE additionally provides diagnostic resolution by explicitly identifying supported and unsupported atomic facts and attributing them to evidence in the source document, enabling

systematic analysis of error patterns that remain hidden in sentence-level evaluations.

Qualitative analysis further illustrates the value of document-grounded, atomic-level evaluation. Table 6.12 shows an example in which a generated entity description is largely correct but contains an incorrect birth date. FENICE correctly validates the supported atomic facts (e.g., nationality and profession) while identifying the unsupported claim (date of birth) and retrieving contradictory evidence from the source document. This behavior would be difficult to diagnose using surface-similarity metrics or reference-based factuality scores alone, yet it is critical when assessing whether automatically generated entries can be trusted for downstream use.

Overall, the observed correlation between FENICE and F-NLI suggests that FENICE captures reliability signals comparable to reference-based evaluation, while providing substantially richer interpretability through claim-level evidence attribution and requiring no gold reference description. These results demonstrate the practical applicability of FENICE as a reference-free, document-grounded evaluation mechanism for automatically generated entity descriptions, supporting informed assessment of generated content prior to its integration into knowledge bases.

Model	Retrieval	AIDA		OOD	
		F-NLI	FENICE	F-NLI	FENICE
T5	✗	22.1	20.4	-8.2	-5.9
T5 _{RAED}	FiD	49.9	37.1	25.6	28.3
Flan-T5	✗	24.8	22.6	-3.7	-1.5
Flan-T5 _{RAED}	FiD	50.2	38.0	25.9	29.1
SmolLM2	✗	27.4	25.1	-5.5	-2.8
SmolLM2 _{RAED}	Prompt	52.9	30.6	21.3	15.4
Llama-3.2	✗	33.8	31.7	-2.0	-0.6
Llama-3.2 _{RAED}	Prompt	55.5	43.2	27.5	31.8

Table 6.11. Factuality evaluation of RAED using sentence-level entailment (F-NLI) and claim-level verification (FENICE) on in-domain (AIDA_{testb}) and out-of-domain (OOD) datasets.

6.6 Summary

Automatically generating and integrating new knowledge is only beneficial if the resulting content is factually accurate, evidence-grounded, and suitable for downstream use. Tradi-

Input Context	<i>[...] The film starring Liam Neeson and <u>Julia Roberts</u> recounts the life of Michael [...]</i>
Generated Description	Julia Roberts is an American actress and filmmaker born on November 17, 1967.
Extracted Atomic Facts	Julia Roberts was born on November 17, 1967. Julia Roberts is American. Julia Roberts is an actress. Julia Roberts is a filmmaker.
FENICE Verification	Not Supported: Julia Roberts was born on November 17, 1967. Supported: Julia Roberts is American. Supported: Julia Roberts is an actress. Supported: Julia Roberts is a filmmaker.
Contradicting Evidence	Julia Fiona Roberts was born on October 28, 1967, at Crawford Long Hospital in Atlanta, Georgia, to Betty Lou Bredemus and Walter Grady Roberts.
Source: <i>Wikipedia Article</i>	

Table 6.12. Qualitative example of FENICE applied to an automatically generated entity description. The model produces a partially correct description containing one unsupported atomic fact (birth date). FENICE decomposes the text into atomic facts, validates each fact against the source document, and provides explicit evidence for the detected inconsistency.

tional evaluation metrics based on surface overlap or semantic similarity to reference texts provide limited insight into factual correctness and are poorly suited to settings where gold descriptions are unavailable, incomplete, or subjective. This chapter addressed the need for reliable and interpretable evaluation mechanisms for validating generated knowledge in knowledge-intensive settings.

We introduced *FENICE*, a document-grounded factuality evaluation framework that extracts atomic claims from generated text and verifies each claim against the source document using natural language inference. This formulation enables fine-grained assessment of factual correctness together with explicit evidence attribution, addressing key limitations of aggregate similarity-based metrics.

Beyond intrinsic evaluation on summarization benchmarks, we demonstrated how FENICE can be applied to assess automatically generated entity descriptions produced by retrieval-augmented models. Since Wikipedia descriptions can be viewed as summaries of

entity pages, we reformulate this setting as a summarization factuality problem by retrieving the text of each entity’s Wikipedia article, removing the entity description section to prevent leakage, and treating the remaining article text as the document and the generated description as the summary. Quantitative analysis showed consistent trends between FENICE and reference-based factuality metrics, while qualitative inspection revealed that claim-level verification provides diagnostic insight into specific factual errors and remaining failure modes. Importantly, FENICE operates without reliance on gold descriptions, enabling evaluation in realistic scenarios where reference summaries are unavailable.

More broadly, this chapter establishes interpretable, document-grounded evaluation as a necessary component of reliable knowledge augmentation pipelines. By enabling systematic validation of generated knowledge prior to integration, such evaluation frameworks support more reliable deployment of automated knowledge maintenance systems. Together with the preceding chapters on representation and dynamic augmentation, this work contributes a coherent framework for advancing reliable knowledge-intensive NLP in evolving environments.

Chapter 7

Conclusions and Future Work

7.1 Overview

This dissertation addressed core challenges in knowledge-intensive Natural Language Processing by advancing three tightly connected dimensions: knowledge coverage and representation, dynamic integration of emerging knowledge, and interpretable validation of generated knowledge. Together, these contributions strengthen the ability of modern AI systems to access broad world knowledge, incorporate new information as it emerges, and assess the reliability of generated content before it is deployed in downstream applications.

A central motivation throughout this work is that many practical and research systems increasingly rely on curated knowledge repositories, such as encyclopedic corpora and structured knowledge bases, as a reference point for grounding inference, supervision, and evaluation, effectively treating them as a proxy for real-world knowledge. At the same time, the underlying sources of this knowledge evolve continuously and at very large scale: new entities emerge, existing entities change, and factual information is constantly updated. As of February 2026, the English Wikipedia includes 7,143,042 articles and averages around 500 new articles per day.¹ In 2024, people from all over the world made 597 million edits to Wikimedia projects; Wikipedia (in all languages) received 180 million edits, and the English Wikipedia alone received 72 million edits.²

This rate of growth and change makes manual maintenance, verification, and in-

¹<https://en.wikipedia.org/wiki/Wikipedia:Statistics>

²<https://en.wikipedia.org/wiki/Wikipedia:Statistics>

tegration increasingly impractical, particularly when knowledge must remain current to support reliable downstream applications. These dynamics motivate the need for automated approaches that (i) broaden knowledge coverage beyond narrow resource and benchmark distributions, (ii) enable systems to incorporate newly available evidence as knowledge evolves to reflect the real state of the world, and (iii) provide validation mechanisms capable of identifying unsupported or incorrect content at a fine level of granularity. In this chapter, we summarize the main contributions of this thesis toward these goals and outline promising directions for future research.

7.2 Summary of Contributions

This dissertation addresses three tightly connected challenges in knowledge-intensive NLP: achieving broad and reliable knowledge coverage in multimodal representations, enabling dynamic integration of emerging knowledge, and establishing interpretable methods for validating generated knowledge. The contributions form a coherent progression that advances these dimensions within a unified perspective spanning representation, augmentation, and evaluation.

Multimodal Knowledge Representation and Coverage. Our first contribution focuses on multimodal knowledge representation and coverage. At the concept level, *Concept-pedia* provides a large-scale semantically annotated multimodal resource designed to expand the diversity of visual and textual concepts available for training and analysis. Building on this resource, the *Concept-10k* benchmark enables systematic evaluation of vision-language models across a wide range of semantic categories that are largely absent from commonly used benchmarks. Unlike existing evaluations that remain heavily concentrated on a narrow subset of concrete object categories, Concept-10k offers a more representative view of real-world conceptual diversity, revealing systematic limitations in current models’ semantic coverage and generalization.

This perspective naturally extends from conceptual abstraction to entity-centric multimodal knowledge. *Enti-pedia* establishes a large-scale resource for pre-training multimodal entity linking models, covering a broad inventory of entities with substantial visual and textual supervision per entity. This design supports learning entity representations that gen-

eralize across datasets and domains, addressing the dataset-specific bias that characterizes most existing multimodal entity linking resources. Together, Concept-pedia, Concept-10k, and Enti-pedia provide complementary coverage across multiple levels of granularity, from concepts to entities, enabling systematic study of multimodal knowledge representation beyond narrow resource and benchmark distributions.

Dynamic Knowledge Augmentation for Emerging Entities. While improved coverage strengthens static representations, real-world knowledge is inherently dynamic, motivating our second contribution. New entities continuously emerge and existing information evolves, making closed-world assumptions and fixed knowledge bases increasingly inadequate. To address this limitation, *RAED* leverages retrieval-augmented generation as a mechanism for dynamic knowledge augmentation. By grounding model predictions in external evidence at inference time, *RAED* enables systems to represent and disambiguate entities that are absent from predefined inventories, reducing dependence on static parametric knowledge. Empirical results demonstrate that this approach not only enables capabilities that prior methods fundamentally cannot support, but also yields strong generalization in out-of-domain settings, where models optimized for specific datasets typically degrade. Quantitative and qualitative analyses, including evaluation with *FENICE*, further indicate that *RAED* produces factually grounded outputs suitable for supporting the continual extension of knowledge bases.

Interpretable Factuality Evaluation. As knowledge is automatically augmented, reliable validation becomes essential, forming the basis of our third contribution. Incorrect or unsupported information, if incorporated into downstream systems, can propagate errors and undermine trust. *FENICE* addresses this challenge by providing interpretable factuality evaluation grounded in atomic facts. Generated text is decomposed into atomic facts, each of which is individually validated against supporting documents using natural language inference. This yields fine-grained verification signals together with explicit evidence attribution, overcoming the limitations of aggregate metrics that obscure the origin of factual errors. By enabling document-grounded validation without reliance on a single gold reference, *FENICE* supports systematic verification of generated knowledge prior to integration into structured repositories.

Ultimately, these contributions advance multimodal knowledge representation across multiple granularities, enable dynamic integration of emerging knowledge through retrieval-augmented generation, and establish interpretable methods for validating generated content. Together, they strengthen the reliability and scalability of knowledge-intensive NLP systems operating in continuously evolving real-world environments.

7.3 Future Work

The work presented in this dissertation advances knowledge representation, augmentation, and evaluation in multimodal settings, providing a foundation for several research directions that remain open. In this section, we outline promising extensions that can further advance the field toward reliable deployment of multimodal models in information-dense real-world applications.

Multimodal Emerging Entities. Emerging entities remain a fundamental challenge for current entity linking systems, with RAED addressing this challenge in the text-only setting. Extending retrieval-augmented knowledge augmentation to multimodal scenarios requires retrieving and integrating visual evidence, such as images associated with an entity, alongside textual sources. This would enable generation and validation of entity descriptions grounded jointly in visual and textual evidence, better reflecting real-world conditions where new entities are introduced through multimodal sources (e.g., illustrated news articles, social media posts, product listings).

A central challenge lies in designing retrieval and fusion mechanisms that determine when visual evidence is necessary, how it should be weighted relative to textual signals, and how to mitigate noise arising from weakly aligned multimodal sources. Moreover, even in the well-established text-only entity linking setting, emerging entities remain underexplored, with TempEL being one of the few publicly available benchmarks that explicitly targets dynamically evolving and newly emerging entities. The multimodal setting further amplifies this challenge, as entities may appear visually before they are explicitly named or documented in text. Developing multimodal benchmarks and resources for emerging entities is therefore important for enabling systematic progress in this direction.

Multilingual and Cross-Cultural Knowledge Coverage. Despite the global nature of knowledge, most wide-coverage resources and evaluation benchmarks remain heavily English-centric. Extending multimodal resources and entity-centric benchmarks to multiple languages would enable more inclusive evaluation, support cross-lingual transfer analyses, and reveal whether models maintain consistent knowledge access across linguistic realizations and cultural contexts.

A promising opportunity arises from the structure of Wikipedia itself, where articles across languages are semantically aligned through interlanguage links and shared entity identifiers. This multilingual connectivity provides a natural foundation for constructing cross-lingual multimodal resources and for studying how conceptual and entity-level knowledge transfers across languages. Leveraging this structure would allow systematic investigation of how models align representations across languages, how cultural salience affects knowledge coverage, and how evidence availability varies across language editions.

Learning from Factuality Signals for Knowledge Generation. Fine-grained factuality signals can be used not only for post-hoc validation, but also to directly improve generation quality. Future work can explore training objectives that incorporate factuality supervision, including reinforcement learning settings where reward functions encourage evidence-grounded correctness and penalize unsupported claims. Such approaches could enable models to internalize stronger notions of factual reliability while maintaining fluency and coverage. This direction requires careful design to ensure training stability, avoid reward hacking, and preserve informative content while improving verifiability.

Temporal Robustness and Continual Knowledge Evaluation. Dynamic knowledge access motivates evaluation protocols that explicitly reflect temporal change. Future benchmarks should measure robustness to knowledge drift, including scenarios where evidence sources evolve, entities change attributes over time, and previously correct information becomes outdated. Developing temporally-aware retrieval strategies, adaptive evaluation frameworks, and longitudinal benchmarks would enable more realistic assessment of knowledge-intensive systems operating in continuously changing environments.

Together, these directions extend the central theme of this dissertation: reliable knowledge-intensive AI requires broad coverage across knowledge granularities, mecha-

nisms for dynamically integrating emerging information, and evaluation methodologies capable of validating knowledge at the level of detail required for dependable downstream use.

Bibliography

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *CoRR*, abs/2404.14219.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. POLYGLOT-NER: massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 586–594. SIAM.

Harith Alani, Zhaochen Guo, and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semant. Web*, 9(4):459–479.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. Smollm2: When smol goes big – data-centric training of a small language model.

Giusepppe Attardi. 2015. Wikiextractor. <https://github.com/attardi/wikiextractor>.

- Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking foundation models with language-model-as-an-examiner. In *Advances in Neural Information Processing Systems*, volume 36, pages 78142–78167. Curran Associates, Inc.
- Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9448–9458.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.
- Lucas Beyer, Olivier J. Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with ImageNet? *CoRR*, abs/2006.07159.
- Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

- Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*, pages 3121–3124.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020a. Evilbert: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 481–487. International Joint Conferences on Artificial Intelligence Organization. Main track.
- Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020b. Fatality killed the cat or: Babelpic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4680–4686. Association for Computational Linguistics.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Ilias Chalkidis and Yova Kementchedjheva. 2023. Retrieval-augmented multi-label text classification. *arXiv preprint arXiv:2305.13058*.
- Shiqi Chen, Siyang Gao, and Junxian He. 2023a. Evaluating factual consistency of summaries with large language models.
- Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. 2023b. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

- Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.
- Gong Cheng, Zhenpeng Li, Xiwen Yao, Lei Guo, and Zhongliang Wei. 2017. Remote sensing image scene classification using bag of convolutional features. *IEEE Geoscience and Remote Sensing Letters*, 14(10):1735–1739.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.
- Simone Conia, Edoardo Barba, Abelardo Carlos Martinez Lorenzo, Pere-Lluís Huguet Cabot, Riccardo Orlando, Luigi Procopio, and Roberto Navigli. 2024. MOSAICo: a multilingual open-text semantically annotated interlinked corpus. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 7990–8004. Association for Computational Linguistics.
- Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023b. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.
- Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Jing Fan, Dennis Aumiller, and Michael Gertz. 2023. Evaluating factual consistency of texts with semantic role labeling. In *Proceedings of the 12th Joint Conference on Lexical and*

- Computational Semantics (*SEM 2023)*, pages 89–100, Toronto, Canada. Association for Computational Linguistics.
- Enrico Fini, Mustafa Shukor, Xiujuan Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. 2025. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9641–9654.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.
- Evgeniy Gabrilovich, Michael Ringgaard, and Amarnag Subramanya. 2013. Facc1: Freebase annotation of clueweb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0).
- Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in*

- Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.
- Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence volume 2*, pages 665–673 (2020).
- Zorik Gekhman, Jonathan Herzig, Roei Aharoni, Chen Elkind, and Idan Szpektor. 2023a. Trueteacher: Learning factual consistency evaluation with large language models.
- Zorik Gekhman, Jonathan Herzig, Roei Aharoni, Chen Elkind, and Idan Szpektor. 2023b. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.
- Karim Ghonim, Andrei Stefan Bejgu, Alberte Fernández-Castro, and Roberto Navigli. 2025a. Concept-pedia: a wide-coverage semantically-annotated multimodal dataset. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34417–34438, Suzhou, China. Association for Computational Linguistics.
- Karim Ghonim, Pere-Lluís Huguet Cabot, Riccardo Orlando, and Roberto Navigli. 2025b. RAED: Retrieval-augmented entity description generation for emerging entity linking and disambiguation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34439–34452, Suzhou, China. Association for Computational Linguistics.
- Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou,

- Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2015. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63. Special Issue on “Deep Learning of Representations”.
- Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.
- Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462, Online. Association for Computational Linguistics.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.
- Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*.
- Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.
- Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. SumPubMed: Summarization dataset of PubMed scientific articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 292–303, Online. Association for Computational Linguistics.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Re-

- trieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.
- Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 8320–8329. IEEE.
- Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.
- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15262–15271. Computer Vision Foundation / IEEE.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIP-Score: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics.

- Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2 : Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. 2018. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8769–8778. Computer Vision Foundation / IEEE Computer Society.
- Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.
- Zhiwei Hu, Víctor Gutiérrez-Basulto, Ru Li, and Jeff Z Pan. 2025. Multi-level matching network for multimodal entity linking. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. I*, pages 508–519.
- Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pages 15696–15707. PMLR.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Nora Kassner, Fabio Petroni, Mikhail Plekhanov, Sebastian Riedel, and Nicola Cancedda. 2022. EDIN: An end-to-end benchmark and pipeline for unknown entity discovery and indexing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8659–8673, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.
- Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of*

- the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.
- Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization.
- Ben Kurtovic. 2013. parserfromhell. <https://github.com/earwig/mwparserfromhell>.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

- Rémi Lebrete, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

- Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*.
- Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023c. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.
- Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023d. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023a. Multi-grained multimodal interaction network for entity linking. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1583–1594.

- Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023b. Chatgpt as a factual inconsistency evaluator for text summarization.
- Giuliano Martinelli, Francesco Molfese, Simone Tedeschi, Alberte Fernández-Castro, and Roberto Navigli. 2024. CNER: Concept and named entity recognition. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8336–8351, Mexico City, Mexico. Association for Computational Linguistics.
- Abelardo Carlos Martinez Lorenzo, Pere-Lluís Huguet Cabot, Karim Ghonim, Lu Xu, Hee-Soo Choi, Alberte Fernández-Castro, and Roberto Navigli. 2024. Mitigating data scarcity in semantic parsing across languages with the multilingual semantic layer and its dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14056–14080, Bangkok, Thailand. Association for Computational Linguistics.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242.
- George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.
- David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

- Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.
- Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4559–4567. [ijcai.org](https://www.ijcai.org).
- Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.
- Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *CoRR*, abs/1910.14599.
- Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2013. Learning multilingual named entity recognition from Wikipedia. *Artif. Intell.*, 194:151–175.

- Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. ReLiK: Retrieve and LinK, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14114–14132, Bangkok, Thailand. Association for Computational Linguistics.
- Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 48–56, Taipei, Taiwan. Association for Computational Linguistics.
- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1946–1958. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2463–2473.
- Luigi Procopio, Simone Conia, Edoardo Barba, and Roberto Navigli. 2023. Entity disambiguation with entity definitions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1297–1303, Dubrovnik, Croatia. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Chenwei Ran, Wei Shen, Jianbo Gao, Yuhan Li, Jianyong Wang, and Yantao Jia. 2022. Learning entity linking features for emerging entities. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7088–7102.
- Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.
- Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.
- Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik. 2021. Imagenet-21k pretraining for the masses. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.
- Mert Bülent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. 2021. Concept generalization in visual representation learning. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9609–9619. IEEE.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35*:

Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessandro Scirè, Andrei Stefan Bejgu, Simone Tedeschi, Karim Ghonim, Federico Martelli, and Roberto Navigli. 2024a. Truth or mirage? towards end-to-end factuality evaluation with llm-oasis. *arXiv preprint arXiv:2411.19655*. To appear in Computational Linguistics.

Alessandro Scirè, Simone Conia, Simone Ciciliano, and Roberto Navigli. 2023. Echoes from alexandria: A large resource for multilingual book summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 853–867, Toronto, Canada. Association for Computational Linguistics.

Alessandro Scirè, Karim Ghonim, and Roberto Navigli. 2024b. FENICE: Factuality evaluation of summarization based on natural language inference and claim extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14148–14161, Bangkok, Thailand. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

- tics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Noam Shazeer and Mitchell Stern. 2018a. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.
- Noam Shazeer and Mitchell Stern. 2018b. Adafactor: Adaptive learning rates with sublinear memory cost. *CoRR*, abs/1804.04235.
- Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization.
- Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.
- Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. Generative multimodal entity linking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7654–7665, Torino, Italia. ELRA and ICCL.
- Alejandro Sierra-Múnera, Linh Le, Gianluca Demartini, and Ralf Krestel. 2024. Melart: A multimodal entity linking dataset for art. *Transactions on Graph Data and Knowledge*, 2(2):8–1.
- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and*

- Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2443–2449. ACM.
- Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE.
- Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. 2023. WikiBio: a semantic resource for the intersectional analysis of biographical events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada. Association for Computational Linguistics.
- Xuhui Sui, Ying Zhang, Yu Zhao, Kehui Song, Baohang Zhou, and Xiaojie Yuan. 2024. MELOV: Multimodal entity linking with optimized visual features in latent space. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 816–826, Bangkok, Thailand. Association for Computational Linguistics.
- Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.
- Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.
- Rohan Taori, Achal Dave, Vaishal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2024.

- Gemini: A family of highly capable multimodal models. arxiv 2023. *arXiv preprint arXiv:2312.11805*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Simone Tedeschi, Valentino Maiorca, Niccol  Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2521–2533. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth e Lacroix, Baptiste Rozi re, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Chen-Tse Tsai, Stephen Mayhew, and Dan Roth. 2016. Cross-lingual named entity recognition via wikification. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 219–228. ACL.
- Ricardo Usbeck, Michael R der, Axel-Cyrille Ngonga Ngomo, Ciro Baron, Andreas Both, Martin Br mmer, Diego Ceccarelli, Marco Cornolti, Didier Cherix, Bernd Eickmann, et al. 2015. Gerbil: general entity annotator benchmarking framework. In *Proceedings of the 24th international conference on World Wide Web*, pages 1133–1143.

- Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Gräsch, Oncel Tuzel, et al. 2025. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19769–19780.
- Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study.
- Junxiong Wang, Ali Mousavi, Omar Attia, Ronak Pradeep, Saloni Potdar, Alexander Rush, Umar Farooq Minhas, and Yunyao Li. 2024. Entity disambiguation via fusion entity decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6524–6536, Mexico City, Mexico. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022b. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 938–948.
- Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022c. WikiDiverse: A multimodal entity linking dataset with diversified

- contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.
- Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023b. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.
- Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.
- Chengmei Yang, Bowei He, Yimeng Wu, Chao Xing, Lianghua He, and Chen Ma. 2023. MMEL: A joint learning framework for multi-mention entity linking. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2411–2421. PMLR.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Klim Zaporozets, Lucie-Aimée Kaffee, Johannes Deleu, Thomas Demeester, Chris Develder, and Isabelle Augenstein. 2022. Tempel: Linking dynamically evolving and newly emerging entities. In *Advances in Neural Information Processing Systems*, volume 35, pages 1850–1866. Curran Associates, Inc.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual*

- Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952. IEEE.
- Pengyu Zhang, Congfeng Cao, Klim Zaporozhets, and Paul Groth. 2024a. Cycle: Cross-year contrastive learning in entity-linking. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3197–3206.
- Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.
- Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022b. EntQA: Entity linking as question answering. In *International Conference on Learning Representations*.
- Zefeng Zhang, Jiawei Sheng, Zhang Chuang, Liangyunzhi Liangyunzhi, Wenyuan Zhang, Siqi Wang, and Tingwen Liu. 2024b. Optimal transport guided correlation assignment for multimodal entity linking. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4103–4117, Bangkok, Thailand. Association for Computational Linguistics.
- Kaiyang Zhou, Jingkan Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt

learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825.