ITA-Bench: Towards a More Comprehensive Evaluation for Italian LLMs

Luca Moroni[1,*,†], Simone Conia[1,†], Federico Martelli[1] and Roberto Navigli[1]

[1]Sapienza NLP Group, Dipartimento di Ingegneria Informatica, Automatica e Gestionale, Sapienza University of Rome, Italy

Abstract

Recent Large Language Models (LLMs) have shown impressive performance in addressing complex aspects of human language. These models have also demonstrated significant capabilities in processing and generating Italian text, achieving state-of-the-art results on current benchmarks for the Italian language. However, the number and quality of such benchmarks is still insufficient. A case in point is the "Open Ita LLM Leaderboard" which only supports three benchmarks, despite being one of the most popular evaluation suite for the evaluation of Italian-language LLMs. In this paper, we analyze the current limitations of existing evaluation suites and propose two ways of addressing this gap: i) a new suite of automatically-translated benchmarks, drawn from the most popular English benchmarks; and ii) the adaptation of existing manual datasets so that they can be used to complement the evaluation of Italian LLMs. We discuss the pros and cons of both approaches, releasing our data to foster further research on the evaluation of Italian-language LLMs.

Keywords

Large Language Models, Natural Language Processing, Evaluation, Italian Language

1. Introduction

LLMs are becoming more and more prominent in NLP, showing impressive results on an increasing range of standard benchmarks, thanks in particular to their reasoning and in-context-learning capabilities [1, 2]. The current trend points towards increasingly larger models trained on massive amounts of data [3, 4]. However, despite these advancements, there remains a significant gap in the availability of high-quality benchmarks for languages other than English, including Italian, which is often considered too optimistically as a high-resource language. Benchmarks are essential for measuring progress in NLP, providing a standardized way to evaluate and compare models, and this is now especially important for Italian given the growing amount of language-specific models that are being developed for the language [5, 6, 7, 8, 9]. High-quality benchmarks must be well-crafted to ensure they accurately reflect the complexities of the language and the specific challenges it presents.

As of today, most of the existing Italian benchmarks are translations of English datasets, which may not fully capture the nuances and unique characteristics of the Italian language. Nevertheless, the ability to automatically translate English benchmarks into Italian is valuable and enticing for two main reasons. First, it provides a way to compare almost 1-to-1 the results obtained in English to the ones obtained in Italian, as the translation process is aimed at keeping an alignment from the source to the target text by design. Second, it provides a quick and relatively simple way of producing a benchmark in Italian, assuming that the translation tool is able to produce high-quality outputs. Unfortunately, the current evaluation suites that are based on automatic translations include only a limited number of benchmarks. For instance, the "Open Ita LLM Leaderboard", which is one of the most popular evaluation suites for Italian LLMs, relies on just three main benchmark translations, namely, MMLU, HellaSwag, and ARC-Challenge. This biases and hampers the assessment, and may not allow the advanced capabilities of modern LLMs to be fully analyzed, even though recent efforts are starting to address this limitation [10].

Having gold LLM benchmarks natively written in Italian is also important, as their scarcity hinders the accurate evaluation of LLMs' capabilities in the Italian language, limiting our understanding of their true performance and potential areas for improvement. Indeed, the translation of English-centric benchmarks may contain instances that refer to concepts, entities, cultures, traditions, historic events, politics, and economics that are not akin to what one is more likely to find in Italian texts and/or in Italy [11, 12, 13]. However, the creation of completely new datasets that take into account such elements is difficult, complex, and time-consuming, and requires expert knowledge. Falling in between automatic translations of existing datasets from English and the creation of brand-new datasets in Italian, there is the option of adapting existing Italian datasets that were originally created for a different purpose, to measure the capabilities of LLMs in Italian language understanding and genera-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.
†These authors contributed equally.
moroni@diag.uniroma1.it (L. Moroni); conia@diag.uniroma1.it (S. Conia); martelli@diag.uniroma1.it (F. Martelli); navigli@diag.uniroma1.it (R. Navigli)

tion. This direction has gained traction over the past few months, with efforts that focus on repurposing Italian tests (usually designed for humans) to evaluate LLMs instead [14].

In this paper, we follow both directions and introduce ITA-Bench, a more comprehensive benchmark suite for the evaluation of Italian-language LLMs. First, ITA-Bench proposes a new extended suite of benchmarks created by automatically translating the most popular English benchmarks into Italian. Second, ITA-Bench includes existing manually curated datasets, adapted to enhance the evaluation framework for Italian LLMs. These two complementary approaches aim to bridge the evaluation gap and provide a more thorough understanding of the capabilities of Italian-language LLMs. With ITA-Bench, we hope to foster further development and refinement of evaluation techniques for Italian LLMs, ultimately contributing to the broader field of multilingual NLP. ITA-Bench is available at https://github.com/sapienzanlp/ita-bench.

2. ITA-Bench: a New Evaluation Suite for Italian LLMs

In this section, we introduce our methodology for the creation of ITA-Bench, a more comprehensive evaluation suite for Italian LLMs. Our objective is to focus on the Italian language and, more specifically, to create a benchmark suite that is able to test a wide variety of aspects of LLMs that "generate" Italian text. To accomplish this objective we focus on two distinct directions: i) translating existing English benchmarks that are currently used to evaluate the capabilities of state-of-the-art LLMs in English, and ii) adapting existing Italian benchmarks, drawing from popular repositories, conferences, shared tasks, and community initiatives, such as the several EVALITA editions[1] and SemEval tasks.[2] In the case of adaptation of existing datasets, most of the work consists in adapting the scope of the tasks, i.e., since many of these tasks were not designed to evaluate LLMs, the core of the work lies in reframing the problem in a way that a prompt can be used to test the capability of a particular LLM to solve a specific task. Table 1 reports the overall statistics of the datasets that we consider for our ITA-Bench suite.

2.1. Translating English Benchmarks

2.1.1. Issues with existing translations

The most popular and widely-used evaluation suite for Italian produced via translation is perhaps the "Open Ita LLM Leaderboard". This is a collection of three datasets – HellaSwag [15], MMLU [16], and ARC-Challenge [17] – that were automatically translated into Italian. Although this set of three benchmarks is generally considered to be of high-quality (thanks to the fact that the translations were produced using GPT-3.5), there are still several issues that limit the quality of this evaluation suite:

Coverage: Open Ita LLM Leaderboard only covers three benchmarks. There are plenty of other datasets that are generally used to test the capabilities of LLMs in English, so limiting the assessment of Italian LLMs to just three datasets may result in the evaluation of some important aspects of their capabilities in Italian being overlooked.

Reproducibility: The code and models used to translate these three benchmarks are not directly available, making it hard – if not impossible – to reproduce the translations.[3]

Transparency: The fact that the translations are not reproducible makes it difficult to analyze whether there are errors or there is margin for improvement in the translation process originally used to translate the three benchmarks.

English specificity: Despite the translation process, these benchmarks actually remain tied to the English language. Indeed, the prompts used as input to the language model contain parts that are in English (for example, in the creation of the examples used for few-shot evaluation). This is undesirable because it inherently favours LLMs that are bilingual, more specifically, LLMs that can "speak" fluent English in addition to Italian.

Uniformity: The translation of benchmarks from English to a target language is usually done on a benchmark-by-benchmark basis. On one hand, this allows developers to specialize the translation code to each dataset; on the other hand, this approach prevents the translation process from being comparable across datasets, which makes performing a root-cause-analysis on the origin of an error in the translated dataset more complex.

2.1.2. Re-translating English benchmarks

Here we describe our methodology that is aimed at addressing the issues that are present in existing benchmark translations, including the ones used in Open Ita LLM Leaderboard. More specifically, we introduce a new library called OBenTO (Open Benchmark Translation for the Others) that is designed to translate existing benchmarks in a uniform, reproducible and fully-transparent way. Moreover, it is also designed to be easily extensible, in such a way that the research community can add new benchmark translations and even

[1]https://www.evalita.it/campaigns/
[2]https://semeval.github.io/

[3]For example, the version of GPT-3.5 used to translate the benchmarks is not known. Also note that OpenAI has already deprecated many GPT 3.5 versions.

new languages besides Italian. We release OBenTO at https://github.com/sapienzanlp/obento.

Translation model. The OBenTO library is designed to be easily adaptable to new backbones, but at the time of writing this article, the library relies on TowerLLM [18], a recent open LLM that is built on top of open-weight LLMs, such as LLaMA and Mistral. TowerLLM continues the pretraining stage on 10 languages to improve multilingual capabilities of the starting LLM. Moreover, TowerLLM is fine-tuned on translation and other translation-related tasks, including grammar error correction, named entity recognition and post-translation correction.

Translated benchmarks. We translate the following datasets from English to Italian:

ARC Challenge and **ARC Easy (ARC-E)** [17, ARC-C, ARC-E]: These are two benchmarks on reasoning and scientific knowledge, created from a single dataset; the ARC Challenge split is obtained by selecting all those questions that QA systems at the time were not able to answer correctly.

GSM8K [19]: A benchmark that tests the capability of an LLM to solve simple math problems whose solution only requires the use of basic arithmetic operations.

BoolQ [20]: A benchmark obtained from queries by search engine users. The task consists in answering Yes or No depending on an input passage that provides context.

HellaSwag [15, HS]: A commonsense reasoning dataset that requires a system to select the most suitable continuation for a given text, based on implicit commonsense knowledge.

MMLU [16]: A benchmark which encompass several questions over 57 subjects across STEM, the humanities, the social sciences, and more.

PIQA [21]: A benchmark that evaluates the capability of an LLM to reason about physical interactions.

SciQ [22]: A reading comprehension test set that challenges an LLM to extract the answer from a passage and question given in input.

TruthfulQA [23, TQA]: A question answering benchmark with a focus on popular misconceptions found across the Web.

Winogrande [24, WG]: a commonsense reasoning dataset that requires choosing between two options based on coreference resolution.

2.2. Adapting Italian Benchmarks

In addition to our new automatically-translated benchmarks, ITA-Bench also includes the adaptation of existing Italian benchmarks from two main sources: the EVALITA campaigns and the SemEval shared tasks. These sources provide Italian data and annotations for a variety of tasks, covering a broad spectrum of linguistic capabilities and phenomena in the Italian language.

The key step in adapting these Italian benchmarks – originally designed for different use cases – is to reframe each task as a question answering task, enabling LLMs to approach and solve them effectively through prompting. In practice, this involves transforming the input of each task into a natural question and the output into a corresponding natural answer or continuation. Where applicable, we also design a set of incorrect answers or distractors of varying complexity. In our adaptation process, we differentiate between two prompting strategies: multiple-choice and cloze style. In the multiple-choice approach, the LLM is given a question along with a predetermined set of possible answers from which it must choose the correct one. In the setting of adapting existing benchmarks, the multiple-choice style will also encompass binary classification prompting, where the only possible responses are "sì" (yes) or "no". In the cloze style approach, instead, the LLM is required to generate the correct answer based solely on the question, or equivalently, the generation of correct class verbalization, for classification tasks. Given the large search space of potential answers in this format, the evaluation focuses on ensuring that the likelihood of the correct answer is higher than that of a predefined set of incorrect answers.

We discuss the details of the adaptation process for each dataset in the following sections and in Appendix C. We offer multiple-choice and cloze style implementations for all datasets except QUANDHO and DISCOTEX, which have only multiple-choice due to their sentence- and paragraph-length choices.

AMI [25]: Automatic Misoginy Identification is a classification task in which the goal is to understand whether or not a tweet is misogynist. The original task is divided into two subtasks, *Behaviour* and *Synth*. Behaviour consists in classifying a tweet into one of three classes, namely, no misogyny, mild misoginy, and aggresive misogyny. Instead, Synth consists of a binary classification task, misogyny v. no misogyny. ITA-Bench includes both subtasks, but in this work we focus on Synth, as Behaviour is more complex due to its unbalanced class distribution.

NERMuD [26]: Named Entity Recognition on Multi-domain Documents was first presented at EVALITA-2023. The task uses standard NER classes, namely, *Person*, *Organization*, and *Location*, to tag entities in a text. In ITA-Bench, we adapt NERMuD and create task instances comprised of three elements: i) the sentence that contains the entity mention, ii) the mention of the entity in the sentence, and iii) the correct class associated with the mention in the given

Dataset	Train set	Valid set	Test set
ARC-C	1068	286	1132
ARC-E	2157	549	2258
GSM8K	7473	-	1319
BoolQ	9399	3259	-
HS	39722	9998	-
MMLU	269	1402	13127
PIQA	15038	1713	-
SciQ	-	983	985
TruthfulQA	-	792	-
Winogrande	4717	1176	-
AMI	7014	-	2908
WiC	2805	500	500
NERMuD	14529	4079	3943
PRELEARN	2328	-	699
PreTENS	5837	-	14560
DISCOTEX	16000	-	1600
GhigliotinAI	62	-	553
QUANDHO	384	-	1416

Table 1
Statistics of the ITA-Bench datasets, for each dataset the cardinalities of the training, validation and test set are reported.

context. We distinguish between two subdomains: *ADG*, writings and speeches from the Italian politician Alcide De Gasperi, and *WN*, news texts from the past decades.

DISCOTEX [27]: Assessing DIScourse COherence in Italian TEXts is a task focused on modelling discourse coherence in real-word Italian texts. In ITA-Bench, we focus only on the first sub-task of DISCOTEX: *Last Sentence Classification*, where, given a short input paragraph and a sentence, the goal is to tell whether the sentence is a valid continuation of the paragraph. To assess the capability of an LLM to solve this task, we reframe DISCOTEX as a multi-choice question answering task. More specifically, given an input paragraph, the LLM is tasked with selecting the most appropriate continuation from among five options that we provide (the original dataset does not provide distractors). Therefore, for the subset of instances with valid continuations, we create a set of distractors by sampling continuations from other instances at random. Instead, for the instances with invalid continuations, we create a new correct option *"nessuna delle precedenti"* (none of the above), and add a set of four random distractors from other instances.

PreTENS[4]: Presupposed Taxonomies was first proposed for SemEval-2022. This task focuses on semantic competence, and evaluates the ability of an LLM to recognize valid taxonomic relationships between two nominal arguments. For example, this can require recognizing whether or not a concept is a subclass of another concept. In ITA-Bench, an LLM is tasked with identifying whether the relationship between two concepts in the same sentence is acceptable.

[4]https://sites.google.com/view/semeval2022-PreTENS

PRELEARN [28]: Prerequisite RElation LEARNing is a task from EVALITA 2020 on concept prerequisite learning. This task consists in identifying whether a concept A is a prerequisite of another concept B, i.e., if learning concept B requires having already learnt concept A. The original dataset comes with four domains, namely, *Geometry*, *Precalculus*, *Physics*, and *Data Mining*, and we maintain these same domains in ITA-Bench.

WiC [29]: Word-in-Context for Italian. We focus on the *binary-classification* sub-task of the original formulation. In ITA-Bench, an LLM is tasked with determining if a word w occurring in two different sentences s_1 and s_2 has the same meaning in s_1 and s_2.

QUANDHO [30]: The QUestion ANswering Data for Italian HistOry dataset is an Italian question-answering dataset focused on Italy's history during the first half of the 20th century. It provides Wikipedia passages that may contain the answer to specific questions. Each question in the dataset appears in multiple (*question*, *answer*) pairs, where the answer can be either correct or incorrect. In ITA-Bench, we select the pair with an answer marked as correct and three distractors from the occurrences of incorrect answers paired with the same question.

GhigliottinAI: Starting from two different EVALITA tasks, nlp4fun [31] and ghigliottin-AI [32], we collect about 600 different games extracted from the TV show and the boardgame of *"L'Eredità"*, a popular quiz game in Italy. *"La Ghigliottina"* is a challenging game that requires extensive knowledge of the Italian culture. The goal is to find a single word that links five seemingly unrelated words. However, since multiple solutions are often possible and computing all potential answers is impractical, in ITA-Bench, we reframe the problem as a multi-choice question answering task, i.e., a simplified version in which four possible words are given and, among these, only one can be linked to all the five input words. In ITA-Bench, we also select three distractor words in such a way that the distractors are linked to three of the five input words. We ensure that the distractors are not too similar one to the other by maximixing the cosine distance of their FastText embeddings. The distractors are also designed to be at most one character shorter or longer than the correct word, resulting in a task that is easy for humans but challenging for LLMs.

3. Evaluation Results

In this section, we discuss the results of various LLMs on ITA-Bench: we first present the results on the automatically-translated benchmarks and then on the adapted benchmarks. ITA-Bench implements all the task formulations using the `lm-evaluation-harness` li-

Type	Size	Name	ARC-C	ARC-E	BoolQ	GSM8K	HS	MMLU	PIQA	SciQ	TQA	WG	AVG
Base	0.4B	Minerva-350M-base-v1.0	24.6	36.4	60.7	48.2	32.6	25.7	59.5	63.7	46.5	58.4	45.6
Base	1B	Minerva-1B-base-v1.0	26.60	42.2	57.1	49.7	39.6	27.0	62.9	73.5	44.6	60.0	48.3
Base	3B	OpenELM-3B	27.0	37.9	60.9	49.7	40.7	28.3	56.7	81.8	47.3	58.4	48.9
Base	3B	XGLM-2.9B	27.5	41.4	59.1	65.7	44.5	27.4	59.9	77.8	43.1	60.2	50.6
Base	3B	Minerva-3B-base-v1.0	31.4	49.1	62.1	55.8	52.9	29.2	66.9	79.9	41.4	62.2	53.1
Base	7B	OLMo-7B-0724-hf	30.7	44.0	72.9	52.5	47.9	30.9	58.7	85.1	44.6	61.2	52.8
Base	7B	LLaMAntino-2-7b	33.7	50.8	70.9	52.2	54.9	33.8	64.4	86.1	44.3	64.1	55.5
Base	7B	Minerva-7B-base-v1.0	38.4	57.7	68.2	52.2	60.4	34.0	69.4	85.2	42.5	63.9	57.2
Base	7B	Mistral-7B-v0.1	42.8	61.3	78.2	56.1	60.4	38.0	65.5	90.8	43.5	68.8	60.5
Base	8B	Llama-3.1-8B	44.0	61.1	78.0	57.8	62.9	38.7	67.7	90.3	43.0	69.2	61.3
Instruct	7B	Mistral-7B-Instruct-v0.1	37.4	55.2	60.4	56.0	52.6	35.7	61.4	85.7	50.8	62.1	55.7
Instruct	7B	Maestrale-chat-v0.4-beta	51.9	71.3	82.9	55.0	69.3	43.7	70.6	92.3	49.6	71.4	65.8
Instruct	8B	LLaMa-3.1-8B-Instruct	49.1	67.2	79.6	61.6	63.5	42.3	67.8	91.4	47.8	69.6	64.0
Instruct	8B	LLaMAntino-3-ANITA	55.9	72.3	76.7	56.9	68.1	46.5	67.0	92.2	57.4	69.9	66.3
Instruct	9B	Italia-9B-Instruct-v0.1	37.1	57.0	62.4	56.6	56.2	32.8	67.8	87.6	38.2	64.0	56.0

Table 2
Evaluation results on standard benchmarks translated to Italian. All LLMs are evaluated using a 0-shot cloze style setting.

brary [33], which allows us to calculate the likelihoods for each possible continuation in a simple and comparable way, as `lm-evaluation-harness` is also used by Hugging Face for the Open LLM Leaderboard.

3.1. Automatic Translation

The results of various LLMs on our translated benchmarks are reported in Table 2, which provides an overview of the zero-shot scores on cloze style task formulations, i.e., the input prompt to an LLM includes only the question without the possible answers. More specifically, we compare the results of several open-weight LLMs having different sizes, ranging from less than 1B parameters up to 9B parameters and focusing on LLMs that have been pretrained, fine-tuned and/or adapted on/to the Italian language. As we can see, the scores of the LLMs are roughly correlated to their size in terms of number of parameters. Notably, the smaller versions of the Minerva LLMs are able to compete with larger models, thanks to the fact that a significant portion of their pretraining dataset is composed of Italian text (rather than English).

3.2. Adapting Italian Datasets

Moving to the adapted benchmarks in ITA-Bench, Table 3 reports the scores of different state-of-the-art models, ranging from 350M parameters models to 9B parameters. Here, we focus on the results of the LLMs in cloze style tasks, except for QUANDHO and DISCOTEX, as ITA-Bench supports only the multi-choice formulation for these two tasks. Unsurprisingly, the size of the LLMs and their pretraining data are discriminators for reaching better results. Most importantly, even the strongest Italian LLMs, such as ANITA, still struggle to compete against their English counterparts. However, as we can see from the results on GhigliottinAI, Italian LLMs seem to perform well and surpass the results obtained by English models. This may indicate that this task needs a different type of competence and/or knowledge in order to be solved. Indeed, we hypothesize that the task requires a deeper understanding of some elements of the Italian culture, e.g., entities and concepts that are more commonly known in Italy than in other countries. Therefore, pretraining and fine-tuning on Italian documents might be the key to obtaining better results in GhigliottinAI.

4. Manual Error Analysis

In order to assess the quality and reliability of our automatically-translated data, we conduct a manual error analysis. To this end, we examine the translations into Italian produced by four language models: two open-source ones, namely, TowerInstruct-7B-v0.2[5] and TowerInstruct-Mistral-7B-v0.2[6] [34], and two proprietary ones, that is, GPT-3.5-turbo and GPT-4o-mini [35].[7] First, we describe the data and the analysis procedure employed. We then discuss the results of our manual analysis and review some crucial error patterns.

4.1. Data and analysis procedure

As the source of the data for our linguistic analysis, we rely on the ARC dataset, which includes multiple-choice question answering in a wide range of domains. Specifically, we randomly select a sample of 100 instances from

[5]https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2
[6]https://huggingface.co/Unbabel/TowerInstruct-Mistral-7B-v0.2
[7]We employ the OBenTO pipeline to process the translations generated by the open-source models. As for GPT-3.5-turbo, we use the translations available at: https://huggingface.co/datasets/alexandrainst/m_arc. We also translate the datasets using GPT-4o-mini with a pipeline similar to the one used for GPT-3.5-turbo.

Type	Size	Model	AMI	GhigliottinAI	NERMuD	PRELEARN	PreTENS	WiC	DISCOTEX	QUANDHO	Avg
-	-	Random Chance	50.00	25.00	33.00	50.00	50.00	50.00	20.00	25.00	33.85
Base	0.4B	Minerva-350M-base-v1.0	50.37	36.34	45.24	47.49	50.72	49.00	18.56	25.49	40.40
Base	1B	Minerva-1B-base-v1.0	50.96	35.44	49.47	52.61	49.88	48.60	17.56	25.98	41.31
Base	3B	XGLM-2.9B	49.86	30.74	54.20	48.25	52.21	48.20	20.63	25.42	41.19
Base	3B	OpenELM-3B	50.17	27.31	69.54	50.25	48.45	50.20	18.69	26.06	42.58
Base	3B	Minerva-3B-base-v1.0	51.47	45.75	58.91	52.61	51.07	48.40	17.37	28.24	44.22
Base	7B	OLMo-7B-0724-hf	55.43	24.23	73.34	50.49	48.75	51.20	40.18	46.18	48.72
Base	7B	LLaMAntino-2-7b	58.91	31.10	85.95	52.86	49.88	50.00	24.63	38.21	48.94
Base	7B	Minerva-7B-base-v1.0	56.15	45.75	81.01	54.87	50.48	48.40	17.87	26.05	47.57
Base	7B	Mistral-7B-v0.1	69.97	40.32	86.04	54.87	60.42	53.20	56.12	72.52	61.68
Base	8B	Llama-3.1-8B	78.02	39.78	88.69	50.12	62.36	55.40	59.43	72.38	63.27
Instruct	7B	Mistral-7B-Instruct-v0.1	69.84	31.28	83.82	54.63	54.99	53.00	44.50	65.25	57.17
Instruct	7B	Maestrale-chat-v0.4-beta	82.60	43.04	89.13	56.88	61.20	60.80	62.69	80.16	67.06
Instruct	8B	LLaMa-3.1-8B-Instruct	85.96	47.92	92.16	51.48	64.76	57.4	65.56	82.76	68.52
Instruct	8B	LLaMAntino-3-ANITA	81.87	48.46	91.94	58.89	62.06	66.8	63.25	73.37	68.33
Instruct	9B	Italia-9B-Instruct-v0.1	53.40	36.17	86.57	53.12	51.33	49.80	24.31	50.78	50.69

Table 3
Few-shot evaluation results on the adapted tasks. We report the results with 5-shot cloze-style prompting, except for DISCOTEX and QUANDHO (light blue), for which we report the results in 2-shot multichoice-style prompting.

the ARC Challenge dataset and we manually analyze the quality of the translations produced by all language models considered. For each instance, we assess the degree of comprehensibility and fidelity of the translation of both questions and answers, assigning a binary label which indicates whether a translation is acceptable or not. Crucially, we distinguish between *minor* and *major* errors depending on the impact on the comprehensibility and fidelity of the target translation. We then identify error patterns, some of which we describe below, highlighting the cases in which the translation impedes understanding of either the questions or the answers, or fails to faithfully reproduce the source text, thus altering the original meaning. Finally, we discuss the results of our analysis. Annotation guidelines are reported in Appendix A.

4.1.1. Key error patterns

As part of our manual annotation process, we identify error patterns, of which we report four key ones, namely: i) *omissions*, which consist in omitting one or multiple source words in the translation; ii) *incorrect terminology*, that is, the incorrect translation of one or multiple terms into the target language; iii) *untranslated source text*, where one or multiple source words are reported as-is in the translation, despite these words not being commonly used in the target language; and iv) *grammatical errors*, which include orthographic, morphological and syntactical errors. Instances of the aforementioned error patterns can be found in Appendix B.

4.1.2. Inter-annotator agreement

In order to assess the reliability of our manual annotations, we compute the inter-annotator agreement. With this aim in view, we select the already-annotated translations produced by one randomly-chosen model and employ a new annotator to assess their quality based on our guidelines. We obtain a Cohen's kappa of 0.85, which indicates a strong agreement.

4.2. Results

Our analysis shows that GPT-4o-mini outperforms all its competitors. With an error rate[8] of 4%, it is markedly more accurate than TowerInstruct-7B-v0.2, which exhibits an error rate of 23%. TowerInstruct-Mistral-7B-v0.2 and GPT-3.5-turbo show a similar performance, that is, 8% and 9% error rate, respectively. Finally, the most frequent error patterns are omissions, especially when considering open-source models, and incorrect terminology.

5. Conclusion

In this paper, we introduce a novel evaluation suite aimed at advancing the Italian community's ability to assess the competencies of LLMs on Italian data. Our approach follows two main directions. First, we define a novel pipeline called OBenTO, which involves translating state-of-the-art English benchmarks into Italian. Second, we rephrase existing Italian benchmarks to be used for prompting and testing large language models. Additionally, we conduct a comprehensive evaluation of the quality of automatically translated benchmarks, highlighting the inherent challenges of such an approach and analyzing the errors made by four LLMs. We hope that our work can provide a solid evaluation framework for evaluating the capabilities of current and future LLMs in Italian.

[8]We emphasize that this error rate does not provide a nuanced evaluation of the aforementioned and other crucial aspects of translation, such as fluency and idiomaticity.

Acknowledgments

Simone Conia gratefully acknowledges the PNRR MUR project PE0000013-FAIR, which fully funds his fellowship. Federico Martelli and Roberto Navigli acknowledge the support of the CREATIVE project (CRoss-modal understanding and gEnerATIon of Visual and tExtual content, Progetti di Interesse Nazionale - PRIN 2020). Finally, we acknowledge the work of the M.Sc. students of Prof. Navigli's multilingual NLP course for the Academic Year 2024, whose contributions provided valuable insights and ideas for the adaptation of existing Italian benchmarks. We acknowledge the CINECA award IsB28_medit under the ISCRA initiative for the availability of high-performance computing resources and support.

References

[1] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, Advances in neural information processing systems 35 (2022) 22199–22213.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[3] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, arXiv preprint arXiv:2203.15556 (2022).

[4] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, Trans. Mach. Learn. Res. 2022 (2022). URL: https://openreview.net/forum?id=yzkSU5zdwD.

[5] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in Italian language, arXiv preprint arXiv:2312.09993 (2023).

[6] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The Italian large language model that will leave you senza parole!, in: F. M. Nardini, N. Tonellotto, G. Faggioli, A. Ferrara (Eds.), Proceedings of the 13th Italian Information Retrieval Workshop (IIR 2023), Pisa, Italy, June 8-9, 2023, volume 3448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 9–17. URL: https://ceur-ws.org/Vol-3448/paper-24.pdf.

[7] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let's push Italian LLM Research Forward!, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4343–4355. URL: https://aclanthology.org/2024.lrec-main.388.

[8] M. Polignano, P. Basile, G. Semeraro, Advanced Natural-based interaction for the Italian language: Llamantino-3-anita, arXiv preprint arXiv:2405.07101 (2024).

[9] R. Orlando, L. Moroni, P.-L. Huguet Cabot, E. Barba, S. Conia, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of Large Language Models trained from scratch on Italian data, Proc. of CLiC-it 2024 – Tenth Italian Conference on Computational Linguistics (2024).

[10] ItaEval and TweetyIta: A new extensive benchmark and efficiency-first language model for Italian, 2024. URL: https://rita-nlp.org/static/ItaEval_TweetyIta_v1.pdf.

[11] D. Hershcovich, S. Frank, H. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. Cabello Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, A. Søgaard, Challenges and strategies in cross-cultural NLP, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6997–7013. URL: https://aclanthology.org/2022.acl-long.482. doi:10.18653/v1/2022.acl-long.482.

[12] R. Navigli, S. Conia, B. Ross, Biases in large language models: Origins, inventory, and discussion, ACM J. Data Inf. Qual. 15 (2023) 10:1–10:21. URL: https://doi.org/10.1145/3597307. doi:10.1145/3597307.

[13] S. Conia, D. Lee, M. Li, U. F. Minhas, S. Potdar, Y. Li, Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024. URL: https://arxiv.org/abs/2410.14057.

[14] A. Esuli, G. Puccetti, The invalsi benchmark: measuring language models mathematical and language understanding in Italian, arXiv preprint arXiv:2403.18697 (2024).

[15] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, Y. Choi, Hellaswag: Can a machine really finish your sentence?, in: A. Korhonen, D. R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL

2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 4791–4800. URL: https://doi.org/10.18653/v1/p19-1472. doi:10.18653/V1/P19-1472.

[16] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=d7KBjmI3GmQ.

[17] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, arXiv preprint arXiv:1803.05457 (2018).

[18] D. M. Alves, J. Pombal, N. M. Guerreiro, P. H. Martins, J. Alves, A. Farajian, B. Peters, R. Rei, P. Fernandes, S. Agrawal, et al., Tower: An open multilingual large language model for translation-related tasks, arXiv preprint arXiv:2402.17733 (2024).

[19] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al., Training verifiers to solve math word problems, 2021, URL https://arxiv.org/abs/2110.14168 (2021).

[20] C. Clark, K. Lee, M. Chang, T. Kwiatkowski, M. Collins, K. Toutanova, Boolq: Exploring the surprising difficulty of natural yes/no questions, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 2924–2936. URL: https://doi.org/10.18653/v1/n19-1300. doi:10.18653/V1/N19-1300.

[21] Y. Bisk, R. Zellers, J. Gao, Y. Choi, et al., Piqa: Reasoning about physical commonsense in natural language, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 7432–7439.

[22] J. Welbl, N. F. Liu, M. Gardner, Crowdsourcing multiple choice science questions, in: L. Derczynski, W. Xu, A. Ritter, T. Baldwin (Eds.), Proceedings of the 3rd Workshop on Noisy User-generated Text, NUT@EMNLP 2017, Copenhagen, Denmark, September 7, 2017, Association for Computational Linguistics, 2017, pp. 94–106. URL: https://doi.org/10.18653/v1/w17-4413. doi:10.18653/V1/W17-4413.

[23] S. Lin, J. Hilton, O. Evans, Truthfulqa: Measuring how models mimic human falsehoods, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 3214–3252. URL: https://doi.org/10.18653/v1/2022.acl-long.229. doi:10.18653/V1/2022.ACL-LONG.229.

[24] K. Sakaguchi, R. L. Bras, C. Bhagavatula, Y. Choi, Winogrande: An adversarial winograd schema challenge at scale, Communications of the ACM 64 (2021) 99–106.

[25] E. Fersini, D. Nozza, P. Rosso, et al., Ami@ evalita2020: Automatic misogyny identification, in: Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), 2020.

[26] A. Palmero Aprosio, T. Paccosi, et al., Nermud at evalita 2023: overview of the named-entities recognition on multi-domain documents task, in: CEUR WORKSHOP PROCEEDINGS, volume 3473, CEUR, 2023.

[27] D. Brunato, D. Colla, F. Dell'Orletta, I. Dini, D. P. Radicioni, A. A. Ravelli, et al., Discotex at evalita 2023: overview of the assessing discourse coherence in Italian texts task, in: CEUR WORKSHOP PROCEEDINGS, volume 3473, CEUR, 2023, pp. 1–8.

[28] C. Alzetta, A. Miaschi, F. Dell'Orletta, F. Koceva, I. Torre, Prelearn@ evalita 2020: Overview of the prerequisite relation learning task for Italian, EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020 (2020) 363.

[29] P. Cassotti, L. Siciliani, L. C. Passaro, M. Gatto, P. Basile, et al., Wic-ita at evalita2023: Overview of the evalita2023 word-in-context for Italian task., EVALITA (2023).

[30] S. Menini, R. Sprugnoli, A. Uva, "who was pietro badoglio?" towards a qa system for Italian history, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 430–435.

[31] P. Basile, M. De Gemmis, L. Siciliani, G. Semeraro, Overview of the evalita 2018 solving language games (nlp4fun) task, EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 75.

[32] P. Basile, M. Lovetere, J. Monti, A. Pascucci, F. Sangati, L. Siciliani, Ghigliottin-ai@ evalita2020: Evaluating artificial players for the language game "la ghigliottina", EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020 (2020) 345.

[33] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h, H. Li, K. McDonell, N. Muennighoff, C. Ociepa, J. Phang, L. Reynolds, H. Schoelkopf, A. Skowron,

L. Sutawika, E. Tang, A. Thite, B. Wang, K. Wang, A. Zou, A framework for few-shot language model evaluation, 2023. URL: https://zenodo.org/records/10256836. doi:10.5281/zenodo.10256836.

[34] D. M. Alves, J. Pombal, N. M. Guerreiro, P. H. Martins, J. Alves, A. Farajian, B. Peters, R. Rei, P. Fernandes, S. Agrawal, P. Colombo, J. G. C. de Souza, A. F. T. Martins, Tower: An open multilingual large language model for translation-related tasks, 2024. arXiv:2402.17733.

[35] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[36] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, W. Macherey, Experts, errors, and context: A large-scale study of human evaluation for machine translation, Transactions of the Association for Computational Linguistics 9 (2021) 1460–1474.

[37] N. Campolungo, F. Martelli, F. Saina, R. Navigli, DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4331–4352. URL: https://aclanthology.org/2022.acl-long.298. doi:10.18653/v1/2022.acl-long.298.

[38] F. Martelli, S. Perrella, N. Campolungo, T. Munda, S. Koeva, C. Tiberius, R. Navigli, DiBiMT: A Gold Evaluation Benchmark for Studying Lexical Ambiguity in Machine Translation, Computational Linguistics (2024) 1–79. URL: https://doi.org/10.1162/coli_a_00541. doi:10.1162/coli_a_00541.

[39] S. Conia, M. Li, D. Lee, U. Minhas, I. Ilyas, Y. Li, Increasing coverage and precision of textual information in multilingual knowledge graphs, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 1612–1634. URL: https://aclanthology.org/2023.emnlp-main.100. doi:10.18653/v1/2023.emnlp-main.100.

A. Annotation Guidelines

In this section, we report the annotation guidelines adopted to ensure consistency throughout our analysis. Annotators receive a document containing the source text and the translations produced by four language models, namely TowerInstruct-7B-v0.2, TowerInstruct-Mistral-7B-v0.2, GPT-3.5-turbo and GPT-4o-mini. Annotators are required to determine the correctness of a translation. In order for a translation to be deemed correct, two key requirements must be satisfied, namely, comprehensibility and fidelity. A translation is considered comprehensible if a native speaker can easily understand the content of both the question and all the answers. Fidelity, on the other hand, refers to the degree to which the translation conforms to the English source text. In order to determine whether both requirements are adequately satisfied, we categorize translation errors as minor or major. While minor errors do not usually hamper the overall comprehensibility and fidelity, major errors - which might even relate to just one single word - significantly impede comprehensibility and fidelity, potentially leading to incorrect interpretations. Based on this categorization, annotators assign a binary label indicating whether the translation is deemed comprehensible and faithful. During the annotation process, annotators are required to identify potential error patterns. Below, we report instances of error patterns often encountered in Machine Translation [36]:

1. **Incorrect translation of source words**: One or more source words are inaccurately translated. This error category also includes the use of incorrect terminology in the translation.

2. **Omission of one or more words**: Words from the source text are missing in the translation.

3. **Incorrect formulation of the output text**: Errors related to the syntactic and semantic structure of the output text.

4. **Untranslated source text**: One or more source words which are reproduced as-is in the output text, despite these words not being commonly used in the target language.

5. **Grammatical errors**: Errors in grammatical agreement, including mismatches in gender and number.

6. **Inadequate register**: The tone or style of the translation does not align with the context of the source text.

7. **Addition of one or more words**: Additional words or phrases (not present in the source text) are included in the translation.

Source	*A chemical called DDT was once used to kill insect pests. When investigations showed this chemical was harmful to some types of birds, the use of DDT stopped. How was the scientific process best able to help scientists understand that DDT was harmful to birds?*
TowerInstruct-7B-v0.2	Un tempo si usava un prodotto chimico chiamato DDT per uccidere gli insetti [...]. Quando le indagini hanno dimostrato che questo prodotto chimico era nocivo per alcuni tipi di uccelli, si è interrotto l'uso del DDT. In che modo il processo scientifico ha potuto aiutare gli scienziati a capire che il DDT era nocivo per gli uccelli?
Source	*The ability to roll the tongue in humans is coded by the dominant allele R. The inability to roll the tongue is coded by the recessive allele r. A man with an RR allele combination for the trait produces a zygote with a woman with an rr allele combination for the trait. Which allele combination could occur in the zygote?*
TowerInstruct-7B-v0.2	[...] Un uomo con una combinazione di alleli RR per il tratto produce uno zigote con una donna con una combinazione di alleli rr per il tratto. Quale combinazione di alleli potrebbe verificarsi nello zigote?

Table 4
Examples of omission. The source text is reported in italics.

B. Examples of key error patterns

In this section, we report examples of the key error patterns described in Section 4.1.1. Specifically, we report instances of *omissions* (Table 4), *incorrect terminology* (Table 5), *untranslated source words* (Table 6) and *grammatical errors* (Table 7). Errors are highlighted by square brackets in red. Importantly, all examples in the aforementioned Tables are considered *major* errors, with the sole exception of the first instance of omission reported in Table 4. Specifically, the omission of the word *pests* has a limited impact on the comprehensibility and fidelity of the translation and, therefore, for the purposes of the task at hand and our analysis, the translation is considered acceptable. As for untranslated source words, we note several issues in the data. As reported in Table 4, we note that GPT-4o-mini translates the term *weathering* as the Italian equivalent of *erosion*. However, *weathering* and *erosion* are two different geological processes. In fact, *weathering* (which could be translated into Italian as *degradazione meteorica*) refers to the breaking down of rocks and minerals at their original location through physical, chemical, or biological means, without the material being moved elsewhere. In contrast, *erosion* involves the removal and transportation of weathered material by agents such as water, wind, or ice. Hence, in translating *weathering* as the Italian equivalent of the word *erosion*, the model fails to capture the precise meaning of the source term, significantly altering the content of the source text. Our error analysis also shows that MT systems still struggle with disambiguation of concepts [37, 38] and entities [39, 13].

C. Adapted Tasks Prompts

In this section we report all the prompts chosen for the adapted tasks. The cloze style prompts are reported in Table 8, while multi-choice-style prompts can be seen in Table 9. For each task we also defined a system prompt, which consists of a text prepended to the model before the sample prompts, the proposed system prompts are reported in Table 10. We present all prompts in the same format as the *LM-Evaluation-Harness* implementation. To ensure clarity and conciseness, we use Jinja templating[9] for all prompts.

D. In-domain Results

PRELEARN and NERMuD have been reported as average accuracies on the main part of this paper. Results are reported in Table 11 and Table 12, looking at each domain separately for the two different tasks. While results for the zero-shot setting are reported in Table 13 and in Table 14. We reported the results twice, dividing the multi-choice and cloze style prompt setting.

E. Other results for adapted tasks

In this section we report other results about adapted tasks. More precisely, in Table 15 are collected the metrics for the adapted tasks in zero shot setting, where all the tasks are proposed in cloze style prompting, except for DISCOTEX and QUANDHO which are reported in multi-choice prompting.

Since we employed a Multi-Choice (MC) style prompting for all adapted tasks. Table 16 presents the results for these tasks in the zero-shot setting, while Table 17 shows the results in the five-shot setting.

[9] https://jinja.palletsprojects.com/en/3.1.x/templates/

Source	*A euglena cell has a structure called an eyespot that detects light. A paramecium does not have an eyespot, and so it cannot detect light. Why doesn't a paramecium need an eyespot?*
TowerInstruct-7B-v0.2	Una cellula **[eugleno]** possiede una struttura chiamata **[macula occhiolare]** che rileva la luce. Un **[parameno]** non possiede una **[macula occhiolare]** e quindi non riesce a rilevare la luce. Perché un **[parameno]** non ha bisogno di una **[macula occhiolare]**?
Source	*A plateau is most likely formed by [a] runoff from a river. [b] weathering by waves. [c] erosion of rock debris. [d] a buildup of cooled lava.*
GPT-4o-mini	Un plateau è più probabilmente formato da [a] deflusso da un fiume. [b] **[erosione]** da onde. [c] erosione di detriti rocciosi. [d] un accumulo di lava raffreddata.

Table 5
Examples of incorrect terminology. The source text is reported in italics. Text within square brackets in black is not present in the source or target text and it has been included for clarity to indicate options.

Source	*The temperature is lower in the valley than on the mountain top.*
TowerInstruct-Mistral-7B-v0.2	**The temperature is lower in the valley than on the mountain top.**
Source	*acquired trait to gain an inherited trait.*
GPT-3.5-turbo	**[trait]** acquisito per guadagnare un trait ereditato.

Table 6
Examples of untranslated source words. The source text is reported in italics.

Source	*A glass is partially filled with water. Five ice cubes are placed in the glass, causing the level of the water to reach the rim of the glass. Which of the following statements best explains the increase in water level?*
TowerInstruct-7B-v0.2	In un bicchiere è **[stato versato]** dell'acqua fino a metà. Sono stati messi cinque cubetti di ghiaccio nel bicchiere, facendo sì che il livello dell'acqua raggiungesse il bordo del bicchiere. Quale delle seguenti affermazioni spiega al meglio l'aumento del livello dell'acqua?
Source	*Which of the following is most likely an adaptation that resulted from habitat destruction?*
GPT-3.5-turbo	Qual è più probabile un**[']**adattamento che è risultato dalla distruzione dell'habitat?

Table 7
Example of grammatical errors. The source text is reported in italics.

AMI	Tweet: "{{text}}". Il tweet (non presenta caratteristiche misogine | presenta caratteristiche misogine).
NERMuD	Data la frase: "{{text}}" L'entità {{target_entity}} è (un luogo | un'organizzazione | una persona)
PreTENS	{{text}} La frase precedente (non ha senso | ha senso)
PRELEARN	{{concept_B}} (non è un prerequisito per {{concept_A}} | è un prerequisito per {{concept_A}})
GhigliottinAI	Date le parole: {{w1}}, {{w2}}, {{w3}}, {{w4}}, {{w5}}. Domanda: Quale tra i seguenti concetti è quello che lega le parole date? {{choice1}} {{choice2}} {{choice3}} {{choice4}} Risposta: ({{choice1}} | {{choice2}} | {{choice3}} | {{choice4}})
WiC	Frase 1: {{sentence1}} Frase 2: {{sentence2}} La parola "{{lemma}}" ha (un significato differente tra le due frasi | lo stesso significato in entrambe le frasi)

Table 8
Cloze-style defined prompts for the adapted tasks.

AMI Synth	Tweet: '{{text}}' Domanda: il tweet presenta caratteristiche misogine? Rispondi sì o no:
NERMuD	Data la frase: "{{text}}" Domanda: A quale tipologia di entità appartiene "{{target_entity}}" nella frase precedente? A. Luogo B. Organizzazione C. Persona Risposta:
PreTENS	{{text}} Domanda: La frase precedente ha senso senso? Rispondi sì o no:
PRELEARN	Domanda: il concetto "{{concept_B}}" è un prerequisito per la comprensione del concetto "{{concept_A}}"? Rispondi sì o no:
QuandHO	Data la domanda: "{{question}}" Quale tra i seguenti paragrafi risponde alla domanda? A. {{choice1}} B. {{choice2}} C. {{choice3}} D. {{choice4}} Risposta:
DISCOTEX	Paragrafo: "{{text}}" Domanda: Quali delle seguenti frasi è la continuazione più probabile del precedente paragrafo? A. "{{choice1}} B. "{{choice2}}" C. "{{choice3}}" D. "{{choice4}}" E. "{{choice5}}" Risposta:
GhigliottinAI	Date le parole: {{w1}}, {{w2}}, {{w3}}, {{w4}}, {{w5}}. Domanda: Quale tra i seguenti concetti è quello che lega le parole date? A. {{choice1}} B. {{choice2}} C. {{choice3}} D. {{choice4}} Risposta:
WiC	Frase 1: {{sentence1}} Frase 2: {{sentence2}} Domanda: La parola "{{lemma}}" ha lo stesso signicato nelle due frasi precedenti? Rispondi sì o no:

Table 9
Multi-Choice-style defined prompts for the adapted tasks.

AMI Synth	Indica se i seguenti tweet presentano caratteristiche misogine.
NERMuD	Data una frase e un'entità, indica se tale entità rappresenta un luogo, un'organizzazione o una persona.
PreTENS	Indica se le seguenti frasi hanno senso.
PRELEARN	Dati due concetti A e B, indica se il primo concetto è un prerequisito per il secondo. Il concetto A è prerequisito per il concetto B, se per comprendere B devi prima aver compreso A. I seguenti concetti appartengono al dominio: {{domain}}.
QuandHO	Ti saranno poste domande di storia italiana. Identifica quali paragrafi contengono la risposta alle domande date.
DISCOTEX	Ti verranno poste delle domande nelle quali è presente un paragrafo, e come possibili risposte varie frasi che possono essere o meno la continuazione del paragrafo. Indica la frase che rappresenta la continuazione più probabile del paragrafo, oppure "nessuna delle precedenti" se nessuna delle continuazioni è corretta.
GhigliottinAI	Ti viene chiesto di risolvere il gioco della ghigliottina. Il gioco della ghigliottina consiste nel trovare un concetto che lega cinque parole date. Tale concetto è esprimibile tramite una singola parola.
WiC	Date due frasi, che contengono un lemma in comune, indica se tale lemma ha lo stesso significato in entrambe le frasi.

Table 10
Description of the tasks used as a system prompt during the evaluation for the adapted tasks.

	PRELEARN				NERMuD	
Model	**Data Mining**	**Geometry**	**Physisic**	**Precalculus**	**AGD**	**WN**
Minerva-350M-base-v1.0	46.46	45.50	51.50	46.50	47.23	42.99
Minerva-1B-base-v1.0	45.45	57.00	52.00	56.00	50.00	48.94
XGLM-2.9B	49.49	45.00	46.50	52.00	46.81	61.59
OpenELM-3B	51.52	47.50	49.50	52.50	67.23	71.85
Minerva-3B-base-v1.0	46.46	52.50	52.50	59.00	57.02	60.81
OLMo-7B-0724-hf	48.48	46.50	52.50	54.50	70.00	76.69
LLaMAntino-2-7b	44.44	53.00	55.50	58.50	83.62	88.28
Minerva-7B-base-v1.0	51.51	50.50	61.00	56.50	78.51	83.51
Mistral-7B-v0.1	50.50	51.50	54.50	63.00	81.70	90.38
Llama-3.1-8B	48.48	47.50	53.00	51.50	87.44	89.95
Mistral-7B-Instruct-v0.1	53.54	55.50	52.50	57.00	80.00	87.64
Maestrale-chat-v0.4-beta	52.53	54.50	60.00	60.50	86.17	92.09
LLaMa-3.1-8B-Instruct	43.43	53.00	55.00	54.50	92.12	92.20
LLaMAntino-3-ANITA	56.56	55.50	63.50	60.00	90.21	93.67
Italia-9B-Instruct-v0.1	49.49	52.00	56.50	54.50	84.47	88.68

Table 11
5-shots results for PRELEARN and NERMuD dataset, separated into different domains. The reported results are obtained evaluating LLMs with cloze style prompting.

	PRELEARN				NERMuD	
Model	**Data Mining**	**Geometry**	**Physisic**	**Precalculus**	**AGD**	**WN**
Minerva-350M-base-v1.0	53.53	47.00	45.00	45.00	35.10	27.37
Minerva-1B-base-v1.0	49.49	50.00	50.00	50.00	42.34	33.23
XGLM-2.9B	52.53	47.00	49.50	49.50	30.43	32.14
OpenELM-3B	48.48	48.00	48.50	46.50	23.83	34.22
Minerva-3B-base-v1.0	48.48	43.50	49.50	46.50	37.44	32.34
OLMo-7B-0724-hf	51.49	49.49	52.50	50.50	87.11	86.38
Minerva-7B-base-v1.0	49.49	49.00	53.00	46.00	26.38	26.82
LLaMAntino-2-7b	51.52	50.00	54.00	55.00	64.89	65.72
Mistral-7B-v0.1	69.69	64.50	66.00	68.50	90.63	92.08
Llama-3.1-8B	59.59	64.50	63.50	61.00	83.40	91.19
Mistral-7B-Instruct-v0.1	55.56	58.50	58.50	61.00	78.72	85.68
Maestrale-chat-v0.4-beta	63.64	71.00	66.50	68.50	92.55	93.76
LLaMa-3.1-8B-Instruct	69.69	72.50	69.50	69.00	90.42	92.69
LLaMAntino-3-ANITA	71.71	74.00	71.50	66.50	90.63	94.08
Italia-9B-Instruct-v0.1	55.56	52.00	56.00	48.00	74.47	82.79

Table 12
5-shots results for PRELEARN and NERMuD dataset, separated into different domains. The reported results are obtained evaluating LLMs with multi-choice prompting.

	PRELEARN				NERMuD	
Model	**Data Mining**	**Geometry**	**Physisic**	**Precalculus**	**AGD**	**WN**
Minerva-350M-base-v1.0	50.51	50.00	50.00	48.00	51.49	54.35
Minerva-1B-base-v1.0	48.48	52.00	60.00	58.00	54.26	67.31
XGLM-2.9B	52.53	46.00	51.50	44.50	48.72	49.75
OpenELM-3B	50.51	43.00	49.00	48.50	35.11	47.16
Minerva-3B-base-v1.0	52.53	51.00	46.50	53.00	71.06	76.67
OLMo-7B-0724-hf	65.66	52.00	52.50	58.50	45.96	55.79
LLaMAntino-2-7b	48.48	47.00	53.00	51.50	50.00	71.01
Minerva-7B-base-v1.0	53.54	50.50	54.50	59.00	47.87	71.61
Mistral-7B-v0.1	60.61	49.00	54.50	53.50	71.91	88.88
Llama-3.1-8B	67.68	41.00	50.50	53.00	88.09	87.41
Mistral-7B-Instruct-v0.1	59.60	52.00	52.00	49.00	50.64	72.37
Maestrale-chat-v0.4-beta	60.61	57.00	54.00	39.00	78.94	82.59
LLaMa-3.1-8B-Instruct	49.49	53.00	55.00	45.50	89.15	90.10
LLaMAntino-3-ANITA	53.54	51.00	51.00	38.50	91.49	93.13
Italia-9B-Instruct-v0.1	60.61	57.50	56.50	55.50	44.89	38.64

Table 13
0-shots results for PRELEARN and NERMuD dataset, separated into different domains. The reported results are obtained evaluating LLMs with cloze style prompting.

	PRELEARN				NERMuD	
Model	**Data Mining**	**Geometry**	**Physisic**	**Precalculus**	**AGD**	**WN**
Minerva-350M-base-v1.0	50.51	50.00	50.00	50.00	20.64	24.98
Minerva-1B-base-v1.0	52.53	46.50	42.00	49.50	20.64	24.81
XGLM-2.9B	47.47	46.00	50.50	48.50	20.64	24.81
OpenELM-3B	50.51	50.00	50.00	50.00	20.64	24.81
Minerva-3B-base-v1.0	50.51	50.00	50.00	50.00	20.64	24.81
OLMo-7B-0724-hf	50.51	49.00	49.00	50.50	65.32	63,96
LLaMAntino-2-7b	49.49	54.50	52.50	51.00	44.68	57.18
Minerva-7B-base-v1.0	50.51	50.00	50.00	50.00	20.64	24.83
Mistral-7B-v0.1	56.57	46.50	49.00	49.00	83.62	88.94
Llama-3.1-8B	54.55	55.00	58.50	51.50	90.00	92.52
Mistral-7B-Instruct-v0.1	49.49	49.00	50.00	49.50	81.91	89.17
Maestrale-chat-v0.4-beta	63.64	59.50	58.00	55.50	90.21	93.36
LLaMa-3.1-8B-Instruct	64.65	53.50	64.50	57.00	90.64	93.30
LLaMAntino-3-ANITA	56.57	51.50	62.00	56.00	90.64	93.50
Italia-9B-Instruct-v0.1	50.51	50.50	51.50	52.00	52.13	64.51

Table 14
0-shot results for PRELEARN and NERMuD dataset, separated into different domains. The reported results are obtained evaluating LLMs with multi-choice prompting.

Model	AMI	GhigliottinAI	NERMuD	PRELEARN	PreTENS	WiC	DISCOTEX	QUANDHO	Avg
Minerva-350M-base-v1.0	47.46	21.52	52.92	49.63	52.93	50.00	18.56	26.41	39.93
Minerva-1B-base-v1.0	50.41	20.80	60.78	54.62	52.93	50.20	17.94	26.84	41.81
XGLM-2.9B	50.45	26.58	49.24	48.63	52.70	50.00	18.81	26.69	40.39
OpenELM-3B	55.47	20.98	41.13	47.75	52.29	50.00	51.19	61.79	47.57
Minerva-3B-base-v1.0	57.60	34.90	73.87	50.76	52.89	50.00	18.50	27.12	45.70
OLMo-7B-0724-hf	51.24	23.15	64.64	49.75	47.06	50.00	25.75	51.69	45.41
LLaMAntino-2-7b	50.55	22.97	60.50	50.00	52.93	50.00	45.94	69.92	50.35
Minerva-7B-base-v1.0	49.69	30.20	59.74	54.38	52.95	50.00	18.81	26.69	42.81
Mistral-7B-v0.1	56.43	28.21	80.40	54.40	46.74	50.00	45.94	69.92	54.00
Llama-3.1-8B	56.57	31.46	87.75	53.04	45.45	50.00	54.63	65.61	55.56
Mistral-7B-Instruct-v0.1	54.47	28.03	61.50	53.15	59.37	50.00	51.20	45.31	50.38
Maestrale-chat-v0.4-beta	65.75	47.74	80.76	52.65	47.31	50.00	23.06	28.32	49.45
LLaMa-3.1-8B-Instruct	86.28	35.44	89.62	50.75	52.18	50.00	66.31	79.38	63.75
LLaMAntino-3-ANITA	50.58	45.21	92.31	48.51	55.95	50.00	62.63	74.36	59.94
Italia-9B-Instruct-v0.1	50.00	30.02	41.77	57.53	52.93	50.00	49.80	29.19	45.15

Table 15
0-shot evaluation results on the adapted tasks, the tasks are proposed in a cloze style prompting, but QUANDHO and DISCOTEX that are proposed in multi-choice style prompting.

Model	AMI	GhigliottinAI	NERMuD	PRELEARN	PreTENS	WiC	Avg
Minerva-350M-base-v1.0	50.48	22.78	22.81	50.13	46.97	48.00	40.20
Minerva-1B-base-v1.0	50.07	24.23	22.72	47.63	47.07	49.40	40.19
XGLM-2.9B	50.17	22.97	22.72	48.12	46.98	48.60	39.93
OpenELM-3B	50.00	23.15	22.72	50.13	47.07	49.80	40.48
Minerva-3B-base-v1.0	50.07	24.95	22.72	50.13	47.07	50.00	40.82
OLMo-7B-0724-hf	50.00	22.24	50.87	57.16	52.94	50.00	47.20
LLaMAntino-2-7b	50.14	29.11	50.93	51.87	47.07	50.80	46.65
Minerva-7B-base-v1.0	50.00	22.60	22.74	50.13	47.07	50.00	40.42
Mistral-7B-v0.1	50.72	40.69	86.28	50.27	47.07	48.80	53.97
Llama-3.1-8B	50.28	38.70	91.26	54.89	47.07	51.40	55.60
Mistral-7B-Instruct-v0.1	62.79	29.11	85.54	49.50	45.56	69.70	57.04
Maestrale-chat-v0.4-beta	62.62	49.19	91.79	59.16	47.16	59.60	61.58
LLaMa-3.1-8B-Instruct	52.72	37.07	91.97	59.91	51.20	50.00	57.15
LLaMAntino-3-ANITA	65.96	38.70	92.07	56.52	52.05	60.40	60.95
Italia-9B-Instruct-v0.1	50.00	24.59	58.32	51.13	47.07	44.63	45.96

Table 16
0-shot evaluation results on the adapted tasks; the tasks are proposed only in a multi-choice style.

Model	AMI	GhigliottinAI	NERMuD	PRELEARN	PreTENS	WiC	Avg
Minerva-350M-base-v1.0	49.20	22.60	31.24	47.63	49.58	50.00	41.70
Minerva-1B-base-v1.0	49.44	25.67	37.78	49.37	51.31	48.20	43.62
XGLM-2.9B	48.35	23.15	31.28	49.63	51.17	44.00	41.26
OpenELM-3B	49.97	26.76	29.02	47.87	49.53	49.20	42.06
Minerva-3B-base-v1.0	48.96	24.95	34.89	46.99	48.72	45.20	41.61
OLMo-7B-0724-hf	60.01	31.65	87.84	53.50	49.64	52.20	55.81
LLaMAntino-2-7b	60.11	25.86	65.31	52.63	52.77	51.00	51.28
Minerva-7B-base-v1.0	53.19	25.85	26.60	49.37	50.72	47.40	42.18
Mistral-7B-v0.1	74.44	43.21	91.36	67.17	54.24	58.00	64.73
Llama-3.1-8B	77.37	49.36	87.29	62.14	65.28	57.60	66.50
Mistral-7B-Instruct-v0.1	68.26	27.67	82.20	58.39	50.10	56.60	57.20
Maestrale-chat-v0.4-beta	84.01	48.10	93.16	67.41	59.88	69.20	70.29
LLaMa-3.1-8B-Instruct	85.72	49.36	91.55	70.17	62.57	65.8	70.86
LLaMAntino-3-ANITA	84.21	45.56	92.35	70.92	63.02	66.20	70.37
Italia-9B-Instruct-v0.1	60.80	28.75	78.63	52.89	43.56	47.40	52.00

Table 17
5-shot evaluation results on the adapted tasks; the tasks are proposed only in a multi-choice style.