

DanteLLM: Let's Push Italian LLM Research Forward! 🍷 🇮🇹

Andrea Bacciu^{§†}, Cesare Campagnano[§], Giovanni Trappolini, Fabrizio Silvestri

Department of Computer, Control and Management Engineering
Sapienza University of Rome

{bacciu, campagnano, trappolini, fsilvestri}@diag.uniroma1.it

Abstract

In recent years, the dominance of Large Language Models (LLMs) in the English language has become evident. However, there remains a pronounced gap in resources and evaluation tools tailored for non-English languages, underscoring a significant disparity in the global AI landscape. This paper seeks to bridge this gap, specifically focusing on the Italian linguistic context. We introduce a novel benchmark, and an open LLM Leaderboard, designed to evaluate LLMs' performance in Italian, providing a rigorous framework for comparative analysis. In our assessment of currently available models, we highlight their respective strengths and limitations against this standard. Crucially, we propose "DanteLLM", a state-of-the-art LLM dedicated to Italian. Our empirical evaluations underscore Dante's superiority, as it emerges as the most performant model on our benchmark, with improvements by up to 6 points. This research not only marks a significant stride in Italian-centric natural language processing but also offers a blueprint for the development and evaluation of LLMs in other languages, championing a more inclusive AI paradigm. Our code at: <https://github.com/RSTLess-research/DanteLLM>

Keywords: Large Language Models (LLMs), Italian LLMs, LLM benchmark leaderboard

1. Introduction

Large language models (LLMs) have seen a rise in recent years, demonstrating great performance across a diverse set of tasks (Radford et al., 2018, 2019; OpenAI, 2023; Touvron et al., 2023a,b; Google, 2023; Jiang et al., 2023). Their capabilities in understanding, generating, and fine-tuning textual information has revolutionized many aspects of machine learning, natural language processing, and even domains beyond (Saharia et al., 2022; Gani et al., 2023; Lian et al., 2023; Tolomei et al., 2023). The majority of these LLMs advancements, however, predominantly concern the English language, leaving a significant linguistic and cultural gap in the global AI landscape. Such scenario limits the global reach and applicability of these models but also undermines the rich nuances of human expression and knowledge available in diverse languages, as stated by Costa-jussà et al. (2022).

The need for LLMs in non-English languages has become increasingly evident, thanks to the development of easy-to-use applications, also non-expert started adopting LLMs such as OpenAI ChatGPT and Google Bard in their everyday work. Yet, the question remains: how should we effectively develop such models? Beyond their creation, how can we ascertain their performance and ensure they meet the desired standards?

Benchmarks play a pivotal role in this regard, serving as standardized tools to evaluate, compare, and

consequently refine these models' performance. Currently, the well-established LLM benchmarks have been developed exclusively for the English Languages, such as: ARC Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021). With very few multilingual exceptions like Winogrande (Emelin and Sennrich, 2021). The lack of diverse benchmarks (1) limits our understanding and optimization of LLMs for a different set of languages (2) underscores a pressing need for more inclusive evaluation tools that reflect the languages spoken worldwide.

In this paper, we focus on filling this gap in the Italian evaluation of LLMs. Therefore, we start by taking inspiration from the established HuggingFace (HF) Leaderboard¹ proposing a novel Italian benchmark. The HF-leaderboard provides a well-recognized ranking for evaluating LLM performance, covering a large amount of domains and text understanding tasks. To evaluate LLMs they use the following tasks: ARC Challenge, HellaSwag, MMLU, TruthfulQA. To compute the metrics and to prompt the LLMs they use `lm-evaluation-harness` tool (Gao et al., 2021; Beeching et al., 2023). Consequently, for our Italian benchmark, we use all the tasks selected by HF-Leaderboard and carry out the evaluation on `lm-evaluation-harness`. To obtain the tasks in Italian, we opt for an automatic translation approach. To evaluate the quality of our translation, we perform two actions: (1) a human evaluation and (2) we leverage ChatGPT-4 as a

[§]Equal contribution.

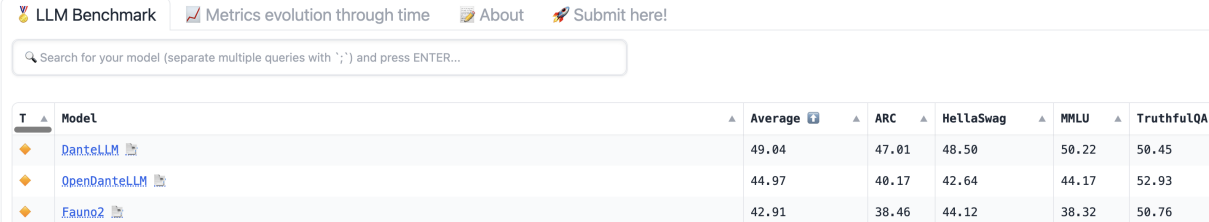
[†]Work done before joining Amazon.

¹https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Open Italian 🇮🇹 LLM Leaderboard 🏆🏆🏆

The 🏆 Open LLM Leaderboard aims to track, rank and evaluate open Italian LLMs and chatbots.

[Eleuther AI Language Model Evaluation Harness](#) - read more details in the "About" page!



T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
👑	DanteLLM	49.04	47.01	48.50	50.22	50.45
👑	OpenDanteLLM	44.97	40.17	42.64	44.17	52.93
👑	Fauno2	42.91	38.46	44.12	38.32	50.76

Figure 1: Italian LLMs Leaderboard demo.

translation evaluator, as done by [Kocmi and Federmann \(2023\)](#); [Leiter et al. \(2023\)](#). In this way, it is possible to present a methodical and reproducible approach to evaluate LLMs in a non-English linguistic context, by laying down a blueprint that can be adapted and refined for other languages.

To address the lack of quantitative assessments of existing models, we use our framework to perform an analysis of the capabilities and limitations of currently available Italian LLMs. Furthermore, we introduce two novel Italian LLMs, DanteLLM and OpenDanteLLM. The former has commercial limitations due to the license of its training data, while the latter has a fully open-source license, even for industrial applications.

Our findings highlight that both DanteLLM and OpenDanteLLM models emerge as the best-performing models on all of our Italian benchmarks. This dual contribution – benchmarks and the Dante models – marks a significant advancement in the realm of Italian-centric natural language processing.

Summing up, the contributions of this paper are as follows:

- Italian-Specific Benchmark: to fill the gap caused by the lack of ways to quantitatively evaluate non-English models, we introduced a novel heterogeneous Italian benchmark for LLMs;
- Comparative Analysis: we are the first to perform and organize a rigorous comparison of existing Italian-based LLMs against this Italian-centric benchmark, bringing light on their respective strengths and weaknesses;
- Online Leaderboard: inspired by the success and usefulness of the HuggingFace LLM Leaderboard, we set up and propose an Italian LLM Leaderboard (Figure 1), to have a public updated assessment over time as new models are made available;

- Introduction of “Dante”(s): two state-of-the-art LLM for Italian, also taking into consideration the importance of licensing in the LLM era for academia and industrial applications;
- blueprint for future research: By establishing a framework for non-English LLM evaluation, this paper paves the way for future research to develop and assess models for other languages, promoting a more inclusive AI ecosystem.

Our models, data, and leaderboard are available at the following link: huggingface.co/rstless-research.

2. Related Work

Large language models (LLMs) have seen a surge in popularity and utility, emerging as a cornerstone in the field of natural language processing (NLP). The initial forays into the realm of neural network-based language models can be traced back to [Benjio et al. \(2000\)](#), who introduced a feed-forward neural network for language modeling. This model paved the way for further exploration and innovation in the space. The introduction of recurrent neural networks (RNNs) ([Medsker and Jain, 2001](#)) marked a paradigm shift in language modeling. [Mikolov et al. \(2010\)](#) leveraged RNNs for language models, showcasing their potential in capturing temporal dependencies. [Hochreiter and Schmidhuber \(1997\)](#) proposed the Long Short-Term Memory architecture, which addressed the vanishing gradient problem inherent in traditional RNNs, enhancing the model’s ability to capture longer-term dependencies. The Transformer architecture, introduced by [Vaswani et al. \(2017\)](#), established a new state-of-the-art in several NLP tasks. This model relied on self-attention mechanisms, dispensing with recurrence and enabling substantial parallelization. [Devlin et al. \(2019\)](#) introduced BERT (Bidirectional Encoder Representations from Transformers), which employs a novel pre-training method using masked language modeling. This

approach, combined with its bidirectional context, facilitated fine-tuning of a myriad of tasks and set new standards in NLP benchmarks (Conia et al., 2021; Campagnano et al., 2022). OpenAI's Generative Pre-trained Transformer (GPT) series serves as a prime example of the capability of LLMs. Starting with GPT (Radford et al., 2018) and extending up to GPT-4 (Brown et al., 2020; Radford et al., 2019; OpenAI, 2023), these models leverage vast amounts of data and billions of parameters. Their unsupervised learning and subsequent fine-tuning on specific tasks have shown unparalleled proficiency in various applications. LLMs have been evaluated and utilized across numerous applications including, but not limited to, text generation, question answering, translation, and summarization (Brown et al., 2020). Notably, LLMs have significantly enhanced the effectiveness of cascading NLP systems (Lewis et al., 2020; Thorne et al., 2021; Trappolini et al., 2023; Bacciu et al., 2023a; Cuconasu et al., 2024). Moreover, the ethical implications, biases in outputs, and the carbon footprint of training such models have also been widely discussed (Hovy and Prabhumoye, 2021; Strubell et al., 2020). Despite their capabilities, LLMs are not without limitations. Issues related to model interpretability, susceptibility to adversarial attacks, and potential amplification of biases have been highlighted in the literature (Wallace et al., 2020; Jia et al., 2020; Trippa et al., 2024).

Italian LLMs The emergence of LLMs has sparked a race to develop models specifically tailored to the Italian language. One of the pioneers was Geppetto (De Mattei et al., 2020), that starting from GPT-2 created a model for the Italian language.

Subsequently, adapters (Houlsby et al., 2019; Hu et al., 2022; Dettmers et al., 2023) have opened up the opportunity of fine-tuning larger models, offering an elegant solution to the challenge of fine-tuning these expansive neural networks. Adapters are a method used to fine-tune pretrained neural networks on new tasks without adjusting the entire model's weights. Instead of retraining the whole network, adapters introduce small, task-specific parameterized modules within the model's layers. These modules are trained to adapt the model to the new task, while the original weights of the LLM remain mostly untouched. This allows for efficient customization and fine-tuning of the model to various tasks without the need for extensive retraining or using large amounts of data.

One of the pioneers in this new wave was Alpaca, as detailed in Taori et al. (2023a), which served as the foundation for the creation of Camoscio (Santilli and Rodolà, 2023). Simultaneously, Bacciu et al. (2023b) introduced Fauno, which currently stands as the largest Italian-based large language model,

built upon the foundation of Baize (Xu et al., 2023; Touvron et al., 2023a,b).

While these models have piqued the interest of the community, they have faced challenges due to the community's relative immaturity. Issues such as bugs trickling down² have been a concern. However, the most pressing need in this field is the development of a robust and comprehensive framework for evaluating these models.

Benchmark for LLMs As the field of natural language processing has advanced, older benchmarks that once challenged state-of-the-art models have become, at least partially, surpassed. These initial datasets and tasks (Marcus et al., 1993; Merity et al., 2016; Wang et al., 2018; Sarlin et al., 2020; Paperno et al., 2016; Petroni et al., 2019), while pioneering, often lack the complexity and diversity to test the capabilities of contemporary large language models. Models quickly saturate these benchmarks, achieving near-perfect scores, rendering them ineffective as measures of true progress. Without evolving benchmarks that reflect real-world challenges and account for a broader range of linguistic nuances, the community risks overfitting to narrow tasks and overlooking areas where improvements are genuinely needed.

As the adoption and complexity of large language models have grown, there's been a discernible need for transparent and easily accessible benchmarks. To address this, the NLP community introduced an open leaderboard for large language models (Beeching et al., 2023; Gao et al., 2021). This leaderboard serves multiple purposes: First, it allows researchers to directly compare state-of-the-art models using a consistent set of metrics and datasets. Second, the open nature of the leaderboard fosters a collaborative environment, encouraging researchers to build upon each other's work. Fourth, the leaderboard covers a wide array of NLP tasks, from traditional ones like text classification and named entity recognition to more complex challenges such as commonsense reasoning and zero-shot learning. Finally, as the field of NLP is highly dynamic, the leaderboard is regularly updated to include new models and results, ensuring it remains a relevant benchmarking tool. However, these benchmarks are presented almost exclusively in the English language. While there have been some attempts to propose evaluations in Italian in the past (Croce et al., 2018), these have been limited to specific tasks and in general, are not thought for the evaluation of LLMs. This severely limits the proper expansion of models thought for other languages. In this paper, we plan to overcome this limitation, with details provided in the next section.

²<https://github.com/huggingface/transformers/issues/22312>

3. Benchmark

As highlighted earlier, we aim to gain a deep insight into the performance and capabilities of modern Italian LLMs. Building on this objective, we contribute with the Italian LLM Leaderboard, taking cues from the widely recognized HuggingFace Leaderboard (HF-Leaderboard). Our benchmarking approach, in fact, uses the same test datasets utilized by the HF-Leaderboard, which includes ARC-C, HellaSwag, MMLU, and TruthfulQA. Furthermore, the HF-Leaderboard streamlines the LLMs evaluation process by providing specialized testing tools, and our approach closely aligns with this, too.

In shaping our benchmarks, we adopted an automated translation technique, opting for the open-source NLLB (Costa-jussà et al., 2022), and in particular, the “facebook/nllb-200-1.3B” version. This model strikes an excellent equilibrium between reproducibility, efficient translation, and high quality. Further post-edit processing was also applied in case of challenging examples (e.g. containing math questions in MMLU). A standout aspect of the NLLB is its compatibility with over 200 languages, marking it as a promising tool for the potential expansion of our framework to encompass a wider range of languages.

The next section clarifies the aspects related to the quality of the translation, while the following ones introduce each of the benchmarks in detail.

3.1. Translation Quality

To assess the translation quality, we sample 100 instances for each tasks, and perform two different quality analyses. The first one consists of letting three humans evaluate the translation quality and then computing their scores; the second one consists of asking ChatGPT-4 to evaluate the translation quality, as shown in (Kocmi and Federmann, 2023; Leiter et al., 2023).

Humans. We use the following guidelines: *Evaluate the English to Italian translation on a scale from 0 to 10 according to these guidelines:*

- *If the translation is ambiguous, assign 0.*
- *If there is a single word that is incorrectly translated and does not change the sentence’s meaning, deduct 2 points. If the incorrect word changes the meaning, assign 0.*
- *For each grammatical error that does not affect the meaning (e.g., wrong article), deduct 0.5 points.*
- *If the translation preserves the original meaning with no errors, assign a score of 10.*

ChatGPT. We use ChatGPT-4 with the OpenAI API to evaluate the translation of our method. We set the model’s temperature to zero for more precise

and stable answers. To instruct ChatGPT to annotate we use the same guidelines used by our human annotators.

Results. We report the results in Table 1, human annotators and ChatGPT-4 report high scores in their evaluation, confirming the reasonable quality of translation produced by NLLB 1.3B. Only HellaSwag reports a lower score due to the more structured composition of the sentences. However, such a score is still reasonable, marking the resource as usable overall;

Metric/Model	ChatGPT 4	Human
ARC Challenge	9.47 ± 1.06	9.45 ± 0.28
HellaSwag	7.95 ± 2.28	7.15 ± 0.62
MMLU	9.13 ± 2.15	9.41 ± 0.23
TruthfulQA	9.20 ± 2.22	9.70 ± 0.22
Avg	8.94 ± 1.93	8.93 ± 0.34

Table 1: Translation quality results with ChatGPT-4 and Human Evaluation.

3.2. ARC Challenge

The AI2 Reasoning Challenge (Clark et al., 2018), commonly known as ARC, was introduced by the Allen Institute for Artificial Intelligence (AI2) as an innovative benchmark to evaluate the reasoning capabilities of machine learning models. Unlike many standard benchmarks, ARC comprises questions extracted from elementary school science exams. While these questions are typically straightforward for humans, they pose significant challenges for machine learning systems. In practice, we use the ARC-C subset of the benchmark, comprising the more demanding set of questions that remain difficult for models, often requiring multi-faceted reasoning or a deeper level of understanding beyond pattern recognition. The dataset is comprised of 2590 questions, of which, 1119 for train, 299 for dev, and 1172 for the test. Testing is commonly performed in a 25-shot fashion.

3.3. HellaSwag

HellaSwag (Zellers et al., 2019), a twist on the phrase “Hella Swag”, is a dataset developed to challenge the capabilities of models in terms of commonsense reasoning. Contrary to traditional question-answering datasets where the task is to pick the right answer from a list, HellaSwag pushes models to reason through ambiguously phrased scenarios to predict the most plausible continuation.

The dataset was crafted using a unique adversarial writing strategy. Initially, a “turker” (a worker from Amazon Mechanical Turk) writes a plausible ending for a given ambiguous prompt. Subsequently, a

language model is used to generate alternative endings. The challenge for other models is to discern the human-written ending from those produced by the language model.

This adversarial approach makes HellaSwag particularly challenging. Given that the distractor endings are generated by a competent language model, they often appear quite plausible. To succeed, models must utilize a deeper level of commonsense reasoning rather than just pattern recognition.

The dataset is comprised of 100 samples for train and 10K for validation. Testing is commonly performed in 10-shot learning.

3.4. MMLU

Measuring Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) dataset emerges as a holistic benchmark designed to assess the performance of models across an extensive range of tasks. Recognizing the multifaceted capabilities of contemporary language models, MMLU challenges these systems not just in specialized domains, but across a broad spectrum of linguistic tasks.

Unlike conventional benchmarks that might emphasize a singular domain of language understanding, MMLU provides an expansive view, aiming to capture the true breadth and depth of a model’s comprehension. By integrating a multitude of tasks (57 categories) into a single evaluation framework, MMLU effectively measures a model’s ability to transition between diverse tasks and grasp varied types of information.

MMLU comprises 15908 questions in total, 1540 of which are part of the validation set and 14079 of the test set. Testing is commonly performed in 5-shot learning.

3.5. TruthfulQA

TruthfulQA: Measuring How Models Mimic Human Falsehoods (Lin et al., 2022), stands out as a critical benchmark designed to evaluate the propensity of LLMs to reproduce or amplify misinformation. Recognizing the potential risks posed by models that inadvertently generate or perpetuate falsehoods, this dataset emphasizes the importance of aligning machine-generated content with factual accuracy. Rather than assessing models on their capability to produce linguistically coherent responses, TruthfulQA probes deeper, examining the veracity of model outputs in the context of questions where humans might typically err or be misled. By deliberately presenting scenarios that toe the line between fact and fiction, TruthfulQA pushes models beyond mere pattern recognition, demanding a discerning grasp of factual information.

The benchmark comprises 817 questions that span 38 categories, including health, law, finance, and

politics. Testing happens in zero-shot since only the test set was provided.

4. DanteLLM

We introduce DanteLLM, a state-of-the-art Italian LLM, built upon the foundation of the Mistral model with 7 billion parameters (Jiang et al., 2023). Mistral not only surpasses the previous open-source state-of-the-art set by LLaMA2 on the same amount of parameters, but achieves higher performance than LLaMA2 with 13 billion parameters. Using Mistral as a foundation model allows to consistently outperform all the previous state-of-the-art Italian LLMs, despite its pre-training on a different language. Further discussion on the performance is detailed in Section 6. We use the following checkpoint: `mistralai/Mistral-7B-Instruct-v0.2`

Hyperparameter	Value
Epochs	3
Learning Rate	$5e^{-5}$
Quantization	8-bit
LoRA R	16
LoRA α	32
LoRA Dropout	0.05
Optimizer	AdamW
Batch Size	128
Checkpointing strategy	min loss

Table 2: Italian benchmark results in zero-shot and standard error. In every benchmark, a higher score means a more accurate answer.

Following Bacciu et al. (2023b) and Santilli and Rodolà (2023), we perform a LoRA fine-tuning with 8-bit quantization of the novel foundation model. To preserve as much as possible the Mistral’s knowledge and to align it to understand and generate fluent Italian text, we apply a fine-tuning of LoRA weights to the attention matrices Q , V and K , and to the projections *gate*, *up* and *down*. We report all the hyper-parameters in Table 2.

We propose two distinct variants of DanteLLM: DanteLLM and OpenDanteLLM, that differ in the training set.

Most of the Italian LLMs (e.g., Fauno and Camoscio) leverage datasets that have been generated with ChatGPT rendering them unsuitable for commercial purposes.

In light of this, we also explored high-quality and fully open-source Italian alternatives to provide the community a truly open-source model. In particular, we leverage the Italian SQuAD dataset (Croce et al., 2018) and 25,000 sentences from the Europarl dataset (Koehn, 2005) (English-to-Italian). The resulting model, OpenDanteLLM, is released under a permissive Apache 2.0 license.

Additionally, we propose DanteLLM, whose objective is to provide a model with the best performances. For that reason, we use the two aforementioned datasets and two of the best Italian LLM train-

ing datasets which are the Fauno’s Quora dataset and the Camoscio dataset.

5. Experimental Setup

In this section, we illustrate our experimental setup.

5.1. Zero- and Few-shot evaluation

The HuggingFace LLM Leaderboard uses a different number of in-context learning examples for each task. The Leaderboard follows this setup: ARC Challenge 25-shot; HellaSwag 10-shot; MMLU 5-shot and only TruthfulQA in zero-shot learning, since the authors proposed a test suite only.

5.2. Quantization

8-bits quantization of model’s parameters (Dettmers et al., 2022) is a powerful technique that allows to reduce the memory footprint of models, with minimal or no sacrifice in performance. Most of the models we evaluate, such as Camoscio and Fauno, are distributed in quantized 8-bit weights format. For these reasons, we use opt for 8-bits quantization in our benchmarks and training. Specifically, to carry out DanteLLM’s training, we use LoRA (Hu et al., 2022). This method allows to fine-tune small rank decomposition matrices, while keeping the rest of the model frozen. Consequently, the model requires significantly less VRAM compared to regular fine-tuning, with minimal to no drop in performance (Xu and McAuley, 2023).

5.3. Hardware Infrastructure

To execute our experiments, we use a machine equipped with a 64-bit CPU Intel i9-10940X, 256GB of RAM, and an NVIDIA RTX A6000 GPU (with 48GB of VRAM) with the OS Ubuntu LTS 20.04.

5.4. Competing Methods

In this section, we introduce all the models that we consider in our Italian LLM leaderboard.

mT5 mT5 (Xue et al., 2021) is a multilingual extension of the T5 v1.1 model (Raffel et al., 2020). The mT5 architecture is an encoder-decoder transformer-based model that comes in five sizes, ranging from a `small` (300 million parameters) up to an `xxl` (13 billion parameters) version. mT5 has been trained on the *multilingual Colossal Clean Crawled Corpus* (mC4) dataset that comprises 101 languages. mC4 contains 186 Million documents, where the Italian portion corresponds to 2.43% of the whole corpus. In our experiments, we use the 3 billion parameters version of the model, which is the closest to the sizes of DanteLLM and OpenDanteLLM. We use checkpoint available at the HF repository `google/mt5-xxl`.

BloomZ (Muennighoff et al., 2022) is a multi-lingual

model trained on 46 languages, including Italian. The authors report that the Italian portion of the dataset only corresponds to the 0.28%. The model comes with different sizes ranging from 560 Million up to 7.1 Billion parameters. To make a fair comparison with the LLaMA1- and LLaMA2-based models, we use BloomZ with 7.1 billion parameters. We use the following HF checkpoint: `bigscience/bloomz-7b1`.

LLaMA2 LLaMA2 (Touvron et al., 2023b) comes with several improvements respect to its predecessor LLaMA (Touvron et al., 2023a). LLaMA2 has been trained with 2.2 trillion tokens (differently from the 1.8 trillion tokens of LLaMA). It uses Grouped-query Attention (Ainslie et al., 2023) and comes in two versions (standard and chat-based) and different sizes, ranging from 7 up to 70 billion parameters. We use the `meta-llama/Llama-2-7b-hf` checkpoint.

Mistral Recently, Jiang et al. (2023) open-sourced Mistral, a new foundation model with 7 billion parameters. It uses Grouped-query Attention, coupled with Sliding Window Attention to perform fast inference. The model has been trained only on English-specific data. Mistral show remarkable results, outperforming the LLaMA2 counterpart with double its parameters. We use the `mistralai/Mistral-7B-Instruct-v0.2` checkpoint.

iT5 After the release of mT5, Sarti and Nissim (2022) presented an Italian-specific version of the T5 v1.1 architecture. The authors applied several pre-processing techniques to the mC4 dataset and kept only the Italian documents. They released three versions of iT5: `small`, `base`, and `large` (ranging from 60 to 738 million parameters, respectively). To compare with our models, we use the largest available model, having the following checkpoint: `gsarti/it5-large`.

GePpeTto (De Mattei et al., 2020) is an Italian model based on the GPT-2 architecture, with 117 millions of parameters. The model has been trained from scratch on two datasets: the ItWac corpus (Baroni et al., 2009) and a dump of the Italian Wikipedia (November 2019) of 2.8GB. Geppetto has been trained for 620.000 steps. We use the following HF checkpoint: `LorenzoDeMattei/GePpeTto`.

Camoscio (Santilli and Rodolà, 2023) Camoscio is a LoRA fine-tuning of LLaMA, with 7 Billion parameters. Camoscio has been trained on an Italian translation of the Alpaca dataset (Taori et al., 2023b). This dataset has been released under a restricted license because it has been generated and translated using OpenAI’s tools. Camoscio applied LoRA fine-tuning only to Q and K matrices of each attention block of LLaMA, architecture for three training epochs. We use the following HuggingFace checkpoint: `teelinsan/camoscio-7b-llama`.

	Model	ARC Challenge	HellaSwag	MMLU	TruthfulQA	Avg
Mult.	mT5	25.94	26.96	25.56	45.50	30.99
	BloomZ	27.30	34.83	36.40	45.52	36.01
Eng.	LLaMA2	33.28	44.31	34.12	44.83	39.14
	Mistral v0.2	37.46	43.48	44.66	54.99	45.15
Italian	iT5	27.39	28.11	23.69	50.49	32.42
	GePpeTto	24.15	26.34	22.87	50.20	30.89
	Camoscio	33.28	42.91	30.53	45.33	38.01
	Fauno1	33.10	43.13	28.79	43.78	37.20
	Fauno2	36.26	44.25	40.30	50.77	42.90
	LLaMAntino	38.22	46.30	33.89	45.03	40.86
Ours	OpenDanteLLM	41.72	46.49	44.25	48.06	45.13
	DanteLLM	41.89	47.99	47.05	52.41	47.34

Table 3: Zero-shot setting. A higher score means a more accurate answer.

	Model	ARC Challenge	HellaSwag	MMLU	TruthfulQA	Avg
Mult.	mT5	27.56	27.86	25.60	–	27.01
	BloomZ	28.24	35.88	31.67	–	31.93
Eng.	LLaMA2	37.71	43.97	37.91	–	39.86
	Mistral v0.2	41.47	42.99	45.84	–	43.43
Italian	iT5	27.99	26.04	24.31	–	26.11
	GePpeTto	25.08	24.99	24.39	–	24.82
	Camoscio	36.60	43.29	29.38	–	36.42
	Fauno1	36.52	42.86	30.45	–	36.61
	Fauno2	39.33	44.07	38.32	–	40.57
	LLaMAntino	41.72	46.91	38.74	–	42.46
Ours	OpenDanteLLM	46.76	46.75	46.89	–	46.80
	DanteLLM	47.01	47.79	48.27	–	47.69

Table 4: Italian benchmark results with few-shot learning following the settings of the HF-Leaderboard. An higher score means a more accurate answer. TruthfulQA benchmark cannot be performed because train and validation examples are not available.

Fauno 1 & 2 (Bacciu et al., 2023b), similarly to Camoscio, are LoRA fine-tunings of LLaMA and LLaMA2, each with 7 billions of parameters. The Fauno family has been trained with the same data as Camoscio, plus a translated conversational dataset generated with ChatGPT-3.5, from the work of Xu et al. (2023). They have applied the fine-tuning to \mathcal{Q} , \mathcal{K} , and \mathcal{V} matrices of each attention block of both architectures, for a single training epoch. We use the following HF checkpoints: `andreabac3/Fauno-Italian-LLM-7B` and `andreabac3/Fauno2-LLaMa2-7B`.

LLaMAntino Basile et al. (2023) released a novel family of LLaMA2-based models, fine-tuned using QLoRA on Italian datasets. They proposed two variants of LLaMAntino – a chat and an instruction-tuned version – both having three different sizes, ranging from 7 up to 70 billion parameters. The chat version has been fine-tuned

with the Italian-translated UltraChat dataset (Ding et al., 2023). The instruction-tuned version, instead, is fine-tuned on the Italian translation of the Dolly dataset (Dolly, 2023) and on the EVALITA 2023 dataset (Caselli et al., 2018). To compare fairly with the other models we rely on the 7 billion variant, using the `swap-uniba/LLaMAntino-2-7b-hf-ITA` checkpoint.

6. Results and Analysis

We construct our Italian benchmark following the established HuggingFace LLM Leaderboard. Notice that the original leaderboard does not report the results of zero-shot experiments for datasets other than TruthfulQA. Instead, for completeness, we also report zero-shot results on all the benchmarks. We test the presented models (see Sections 5.4, 4) in two settings: zero-shot (in Table 3) and few-shot (in Table 4).

6.1. Competing Models

GePpeTto: As can be noticed, on both zero- and few-shot settings GePpeTto obtains the lowest score of the leaderboard, except in TruthfulQA. The explanation for this result lies in the task, which is closer to its training objective since it has been trained on causal language modeling, rather than question answering or instructions.

Regarding the other tasks, GePpeTto’s limitations are related to its limited knowledge (its training set) and the small amount of parameters, compared to the other models. This can also be seen in Table 5, the model struggles especially in some categories, such as statistics, security study, and astronomy.

T5-based models: T5-based models do not excel in these benchmarks, in both zero- and few-shot scenarios. Nevertheless, `iT5-large` has been trained on less data and is a smaller model compared with `mT5-x1`, `iT5` achieves similar performance to its multilingual counterpart, thanks to its fine-tuning on Italian data.

BloomZ: BloomZ, on the other hand, has been trained on a large amount of data with a cross-lingual training objective, has a broader knowledge, and can reuse concepts learned across multiple languages. In fact, despite the limited amount of Italian text in its training corpus, it is able to respond with good accuracy in most of the benchmarks.

Camoscio & Fauno1: Camoscio and Fauno1 are similar models, both are a LoRA fine-tuning of LLaMA1, as mentioned in Section 5.4. In fact, the performance of both models, on average, differ by less than 0.6%. There is a small difference that does not show an absolute winner between the two models: Camoscio shows a slightly better performance in zero-shot, while Fauno1 obtains a higher result in the few-shot setting. Their approach shows a good trade-off between accuracy and a limited computing budget.

The major limitation of their approaches lies in their tokenizers, since these were not trained on Italian text, thus forcing the LoRA weights to learn how to map a sequence of English subwords into an embedding useful to generate Italian sentences.

Fauno2, LLaMAntino & LLaMA2: Fauno2 shares the same fine-tuning method and data as Fauno1; the difference lies only in the underlying model. Indeed, it uses LLaMA2 as pre-trained foundation model, and, therefore, it inherits all the limitations of the original Fauno1. The same situation happens with LLaMAntino, another QLoRA fine-tuned of LLaMA2 on Italian data. From our results, the performance of Fauno2 is higher than those of LLaMAntino, considering both zero- and few-shot settings. Thanks to the fine-tuning on Italian data, Fauno2 and LLaMAntino perform better than their foundation model, LLaMA2.

Mistral Surprisingly, Mistral, without any direct

Italian-specific supervision, can match the performance of Italian-tuned models such as Fauno2 and LLaMAntino. This is probably due to high-quality training data, combined with some contamination of Italian language in its pre-training corpora.

6.2. DanteLLM and OpenDanteLLM

As mentioned in Section 4, one of the objectives of this work is to propose a reproducible, open-source state-of-art model in Italian. In our experiments, DanteLLM and OpenDanteLLM outperform all the currently available Italian LLMs of the same size. In particular, in zero-shot their best competitor is Fauno2, DanteLLM improves by $\sim 3\%$, and OpenDanteLLM improves by $\sim 2\%$, on average.

In the few-shot setting, their best competitor is LLaMAntino, where DanteLLM brings an improvement of $\sim 6\%$ and OpenDanteLLM $\sim 4\%$, on average.

This is also in line with the findings of the respective foundation models: on most benchmarks, Mistral outperforms LLaMA2.

With respect to its foundation model (Mistral), DanteLLM improves by $\sim 5\%$ in ARC Challenge, $\sim 4\%$ in HellaSwag, $\sim 2\%$ in MMLU, and a performance reduction of $\sim 2\%$ in TruthfulQA. The latter is probably due to the nature of the TruthfulQA benchmark: since it is a sentence completion benchmark, Mistral better preserves its original training objective. On average DanteLLM improves by $\sim 3\%$ the overall performance of Mistral.

7. Conclusion and Future Work

In this work, we present a novel Italian evaluation resource, following the lead of the benchmarks used in the well-known HuggingFace LLM Leaderboard. Such a resource is essential for evaluating the capabilities of models in various tasks and domains in Italian, which suffers from poor coverage.

Along with this, we evaluate the performance of all currently available Italian autoregressive LLMs. We set up an online leaderboard, which will be updated as new Italian or multilingual models are made available.

Finally, we propose two novel fully Italian LLMs, namely DanteLLM and OpenDanteLLM, which outperform their counterparts by up to 6 points in our benchmarks. The latter, in particular, is released under the Apache 2.0 license (completely open source, even for commercial use), to bridge the gap between state-of-the-art research and industrial applications in the Italian landscape.

We present this contribution in the hope that it promotes further research in non-English languages, especially those with limited resources. This marks our contribution to Italian LLM research.

For future work, we plan to extend this research to more languages, especially low-resource ones, and provide robust baselines inspired by DanteLLM and OpenDanteLLM.

Category	IT5	Geppetto	Camoscio	Fauno1	Fauno2	LLaMAntino	OpenDanteLLM	DanteLLM
Abstract algebra	21.00 / 22.00	20.00 / 27.00	27.00 / 32.00	25.00 / 30.00	28.00 / 31.00	31.00 / 30.00	30.00 / 30.00	36.00 / 34.00
Anatomy	19.26 / 19.26	19.26 / 22.96	27.41 / 28.89	33.33 / 25.19	36.30 / 40.00	33.33 / 40.74	36.30 / 41.48	42.22 / 37.78
Astronomy	17.11 / 17.76	17.11 / 17.76	27.63 / 27.63	25.66 / 24.34	39.47 / 29.61	32.89 / 40.79	50.66 / 53.29	55.26 / 54.61
Business ethics	30.00 / 28.00	27.00 / 29.00	29.00 / 29.00	32.00 / 31.00	34.00 / 31.00	34.00 / 34.00	44.00 / 46.00	47.00 / 50.00
Clinical knowledge	23.02 / 21.51	20.75 / 21.51	32.83 / 28.30	30.94 / 24.91	40.75 / 38.49	36.60 / 41.51	49.43 / 53.96	53.58 / 53.58
C. biology	26.39 / 26.39	25.00 / 25.69	24.31 / 32.64	29.17 / 35.42	36.81 / 30.56	35.42 / 38.19	42.36 / 41.67	42.36 / 43.75
C. chemistry	18.00 / 20.00	19.00 / 19.00	33.00 / 24.00	25.00 / 27.00	28.00 / 33.00	26.00 / 36.00	37.00 / 34.00	38.00 / 38.00
C. computer science	27.00 / 40.00	24.00 / 26.00	27.00 / 27.00	20.00 / 31.00	36.00 / 43.00	22.00 / 33.00	42.00 / 46.00	40.00 / 42.00
C. mathematics	26.00 / 21.00	19.00 / 21.00	29.00 / 28.00	31.00 / 23.00	32.00 / 30.00	27.00 / 33.00	32.00 / 29.00	35.00 / 30.00
C. medicine	21.97 / 20.81	20.23 / 20.81	31.79 / 20.81	27.17 / 25.43	31.79 / 32.37	29.48 / 30.64	41.04 / 43.93	42.77 / 42.20
C. physics	23.53 / 22.55	20.59 / 21.57	26.47 / 24.51	26.47 / 23.53	26.47 / 20.59	21.57 / 24.51	31.37 / 28.43	36.27 / 29.41
Computer security	27.00 / 19.00	29.00 / 28.00	22.00 / 30.00	24.00 / 35.00	46.00 / 48.00	32.00 / 42.00	50.00 / 53.00	54.00 / 54.00
Conceptual physics	25.96 / 25.53	26.38 / 26.38	32.77 / 33.62	34.89 / 34.04	34.47 / 37.45	36.17 / 35.32	41.28 / 44.26	42.55 / 47.23
Econometrics	23.68 / 23.68	24.56 / 23.68	22.81 / 28.07	19.30 / 24.56	28.95 / 26.32	25.44 / 28.95	25.44 / 33.33	33.33 / 31.58
Electrical engineering	24.83 / 23.45	24.14 / 24.14	27.59 / 27.59	26.21 / 29.66	42.76 / 40.69	40.69 / 44.83	52.41 / 54.48	52.41 / 51.72
Elementary math.	20.90 / 21.96	21.43 / 20.90	27.78 / 21.43	27.25 / 22.49	28.84 / 30.95	26.46 / 27.25	31.48 / 31.75	32.28 / 31.22
Formal logic	32.54 / 29.37	23.02 / 29.37	30.95 / 27.78	23.02 / 34.13	26.98 / 25.40	29.37 / 26.19	31.75 / 29.37	22.22 / 29.37
Global facts	19.00 / 18.00	18.00 / 18.00	29.00 / 29.00	31.00 / 32.00	34.00 / 37.00	38.00 / 38.00	33.00 / 33.00	29.00 / 30.00
H. S. biology	19.35 / 20.97	19.35 / 17.74	33.55 / 28.06	28.71 / 26.13	46.77 / 40.32	35.81 / 42.58	47.42 / 52.58	52.26 / 57.42
H. S. chemistry	19.70 / 16.26	15.27 / 29.56	31.53 / 25.12	24.63 / 21.67	27.09 / 35.96	32.02 / 27.09	35.96 / 42.36	39.90 / 38.92
H. S. computer science	28.00 / 33.00	25.00 / 30.00	20.00 / 25.00	20.00 / 33.00	33.00 / 35.00	32.00 / 37.00	42.00 / 48.00	45.00 / 52.00
H. S. european history	25.45 / 23.03	21.82 / 26.67	35.15 / 30.30	33.33 / 30.30	46.06 / 31.52	29.70 / 33.94	49.70 / 50.30	50.30 / 52.12
H. S. geography	18.69 / 17.68	18.18 / 17.68	33.84 / 29.80	23.23 / 31.82	44.44 / 44.44	38.38 / 40.91	50.51 / 56.57	54.04 / 61.11
H. S. gov.t and politics	21.24 / 19.17	19.17 / 19.69	30.05 / 27.98	21.76 / 25.91	48.70 / 42.49	32.12 / 42.49	51.30 / 58.55	58.03 / 64.25
H. S. macroeconomics	22.31 / 20.51	20.51 / 20.26	28.46 / 25.38	23.08 / 28.46	38.21 / 37.44	32.82 / 40.00	40.77 / 41.28	43.33 / 41.28
H. S. mathematics	22.22 / 23.33	22.59 / 27.04	25.93 / 23.70	27.41 / 24.81	26.67 / 30.37	29.63 / 27.78	28.52 / 30.37	31.11 / 29.63
H. S. microeconomics	19.75 / 21.01	21.01 / 21.01	30.67 / 28.15	25.63 / 30.25	40.34 / 38.66	31.51 / 36.97	49.16 / 45.38	45.80 / 50.42
H. S. physics	21.19 / 19.87	20.53 / 33.11	27.81 / 28.48	24.50 / 29.14	27.15 / 32.45	29.80 / 34.44	28.48 / 29.14	27.81 / 30.46
H. S. psychology	19.27 / 25.14	19.45 / 34.86	37.43 / 32.66	27.71 / 31.74	52.11 / 41.83	41.10 / 43.85	52.29 / 55.41	57.43 / 58.17
H. S. statistics	25.93 / 15.28	15.74 / 15.28	32.41 / 32.41	20.37 / 40.28	27.78 / 27.31	20.37 / 44.44	36.11 / 37.04	30.09 / 37.96
H. S. us history	23.53 / 27.45	25.49 / 24.02	30.88 / 28.43	30.39 / 32.35	44.12 / 33.33	30.39 / 29.90	45.10 / 53.43	48.53 / 48.53
H. S. world history	25.74 / 27.00	27.43 / 25.74	32.91 / 30.38	30.80 / 32.91	50.21 / 41.35	32.91 / 37.55	41.35 / 48.95	47.68 / 49.79
Human aging	30.49 / 31.39	31.39 / 20.18	31.39 / 39.91	39.01 / 39.46	44.84 / 41.70	33.18 / 42.15	48.88 / 54.26	52.91 / 56.05
Human sexuality	23.66 / 25.95	25.95 / 25.95	32.82 / 32.82	29.01 / 30.53	45.04 / 41.98	45.80 / 48.85	56.49 / 58.02	57.25 / 61.07
International law	23.97 / 29.75	23.97 / 23.97	40.50 / 33.88	38.84 / 30.58	61.98 / 54.55	40.50 / 63.64	48.76 / 61.16	57.02 / 60.33
Jurisprudence	21.30 / 25.93	26.85 / 25.93	29.63 / 35.19	41.67 / 37.96	45.37 / 45.37	35.19 / 44.44	49.07 / 55.56	65.74 / 57.41
Logical fallacies	20.86 / 25.15	22.09 / 25.77	33.13 / 25.15	28.22 / 31.29	42.94 / 37.42	34.36 / 37.42	49.69 / 51.53	54.60 / 54.60
Machine learning	30.36 / 30.36	32.14 / 31.25	20.54 / 24.11	25.89 / 28.57	35.71 / 35.71	27.68 / 26.79	43.75 / 38.39	35.71 / 41.96
Management	17.48 / 18.45	17.48 / 37.86	33.98 / 22.33	30.10 / 33.01	48.54 / 45.63	33.98 / 39.81	61.17 / 66.99	66.99 / 66.02
Marketing	29.06 / 26.07	28.63 / 29.06	36.75 / 30.77	34.19 / 36.32	64.53 / 55.98	43.59 / 51.28	62.82 / 67.52	69.66 / 71.79
Medical genetics	25.00 / 30.00	29.00 / 20.00	27.00 / 35.00	35.00 / 28.00	45.00 / 40.00	39.00 / 44.00	44.00 / 52.00	50.00 / 54.00
Miscellaneous	23.63 / 23.75	23.63 / 28.74	38.83 / 36.91	37.04 / 33.84	56.32 / 51.47	46.74 / 50.83	55.68 / 59.51	62.96 / 60.41
Moral disputes	23.12 / 26.59	24.57 / 24.86	30.06 / 27.46	34.97 / 30.64	48.84 / 45.66	35.84 / 44.22	48.84 / 51.16	52.60 / 53.76
Moral scenarios	23.80 / 23.80	23.91 / 23.80	24.13 / 23.58	26.48 / 24.25	23.24 / 26.03	24.25 / 25.14	24.25 / 25.47	27.15 / 24.47
Nutrition	21.90 / 22.55	22.22 / 22.55	34.97 / 33.01	32.35 / 35.62	46.73 / 42.48	40.52 / 40.20	56.54 / 57.19	53.92 / 55.88
Philosophy	18.33 / 27.01	18.65 / 29.90	33.44 / 30.87	33.12 / 32.48	47.27 / 45.34	38.91 / 50.80	47.59 / 54.34	53.38 / 58.52
Prehistory	20.06 / 20.37	20.37 / 21.60	30.56 / 28.09	29.01 / 26.54	40.12 / 38.89	36.73 / 41.05	48.77 / 51.85	50.31 / 56.79
P. accounting	25.53 / 24.47	22.70 / 23.40	27.30 / 21.63	24.82 / 24.82	28.37 / 29.43	28.37 / 32.62	31.56 / 28.72	32.27 / 32.27
P. law	25.49 / 25.42	23.53 / 25.16	25.10 / 25.95	26.53 / 27.18	28.68 / 29.14	25.75 / 28.49	30.38 / 32.86	32.14 / 33.96
P. medicine	27.94 / 43.01	18.38 / 18.38	41.18 / 40.81	22.43 / 43.01	30.88 / 33.82	35.66 / 40.44	35.29 / 37.13	37.87 / 40.07
P. psychology	24.35 / 25.00	25.16 / 25.00	27.78 / 26.14	28.92 / 28.59	37.75 / 33.82	30.88 / 31.21	39.22 / 41.67	42.32 / 45.75
Public relations	21.82 / 21.82	21.82 / 21.82	33.64 / 29.09	32.73 / 30.00	45.45 / 44.55	41.82 / 41.82	43.64 / 52.73	53.64 / 56.36
Security studies	20.00 / 23.67	18.78 / 18.37	32.24 / 31.02	19.59 / 31.43	48.98 / 46.94	29.39 / 45.71	52.65 / 58.37	57.55 / 55.10
Sociology	22.89 / 24.38	24.38 / 23.88	42.79 / 29.35	38.31 / 35.82	53.73 / 47.26	46.27 / 52.24	66.67 / 69.15	67.66 / 70.15
Us foreign policy	28.00 / 28.00	28.00 / 28.00	30.00 / 46.00	30.00 / 39.00	61.00 / 57.00	45.00 / 53.00	61.00 / 67.00	68.00 / 69.00
Virology	30.72 / 27.71	27.71 / 28.31	31.33 / 30.72	34.94 / 27.11	40.96 / 36.14	34.94 / 33.13	45.18 / 43.37	45.78 / 41.57
World religions	30.99 / 25.15	32.16 / 21.05	32.16 / 38.60	35.09 / 38.01	60.82 / 60.23	51.46 / 54.97	70.18 / 69.59	69.01 / 71.35
Avg.	23.69 / 24.31	22.87 / 24.39	30.53 / 29.38	28.79 / 30.45	40.30 / 38.32	33.89 / 38.74	44.25 / 46.89	47.05 / 48.27

Table 5: Detailed results on the MMLU Results (zero-shot / few-shot), “H. S.,” “P” and “C” stand for High School, Professional and College, respectively. Due to space constraints, only Italian models are reported. Our proposed models are highlighted in green.

Limitations

In this section, we report the major limitations of our study. Our benchmark is built on top of an automatic neural translation process, which, despite the overall good translation performances that we evaluated with humans and ChatGPT-4 in Section 3.1, hardly competes with a manually-curated resources. However, our approach can be seen as a starting point in the direction of higher-quality resources. Another limitation consists in the LoRA fine-tuning on an existing model, rather than a training from scratch. A more direct alignment of tokenizer wordpieces and model parameters to the

Italian language would be ideal. However, this approach shows anyways remarkable performances and efficiency.

Ethics Statement

The human annotators were paid 8.47\$ per hour, which is above the minimum wage. To run our training, we produced a negligible amount of CO₂ due to the Parameter Efficient Fine-Tuning techniques employed to train our DanteLLM models. Furthermore, to speed up and obtain a more efficient inference, we applied an 8-bit quantization to all the models in our benchmarks. The resources that

we proposed are translated automatically using the open-source model NLLB 1.3B. Despite the good translation quality reported by human annotators and ChatGPT-4, the NLLB model can still hallucinate and induce some bias in the benchmark data. The same problems are reflected in our proposed LLMs, since there are no guarantees on the safety of their answers.

Acknowledgment

This project was supported by the projects FAIR (PE0000013), SERICS (PE0000014), and IR0000013-SoBigData.it PNRR and the NEREO (Neural Reasoning over Open Data) project PRIN Grant no. 2022AEFHAZ.

References

- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. [Gqa: Training generalized multi-query transformer models from multi-head checkpoints](#).
- Andrea Bacciu, Florin Cuconasu, Federico Siciliano, Fabrizio Silvestri, Nicola Tonellotto, and Giovanni Trappolini. 2023a. [Rraml: Reinforced retrieval augmented machine learning](#). In *Proceedings of the Discussion Papers - 22nd International Conference of the Italian Association for Artificial Intelligence (AIXIA 2023 DP) co-located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIXIA 2023), Rome, Italy, November 6-9, 2023*, volume 3537 of *CEUR Workshop Proceedings*, pages 29–37. CEUR-WS.org.
- Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. 2023b. [Fauno: The italian large language model that will leave you senza parole!](#)
- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43:209–226.
- Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. [Llamantino: Llama 2 models for effective text generation in italian language](#).
- Edward Beeching, Clémentine Fourier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Cesare Campagnano, Simone Conia, and Roberto Navigli. 2022. [SRL4E – Semantic Role Labeling for Emotions: A unified evaluation framework](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Dublin, Ireland. Association for Computational Linguistics.
- Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Sixth evaluation campaign of natural language processing and speech tools for italian: Final workshop (evalita 2018). In *EVALITA 2018*. CEUR Workshop Proceedings (CEUR-WS.org).
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#).
- Simone Conia, Andrea Bacciu, and Roberto Navigli. 2021. [Unifying cross-lingual semantic role labeling with heterogeneous linguistic resources](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 338–351, Online. Association for Computational Linguistics.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Meja Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#).

- Daniilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in Italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. [The power of noise: Redefining retrieval for rag systems](#).
- Lorenzo De Mattei, Michele Cafagna, Felice Dell’Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves Italian into a language model. *arXiv preprint arXiv:2004.14253*.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. [Llm.int8\(\): 8-bit matrix multiplication for transformers at scale](#).
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.
- Free Dolly. 2023. Introducing the world’s first truly open instruction-tuned LLM. databricks.com.
- Denis Emelin and Rico Sennrich. 2021. [WinoX: Multilingual Winograd schemas for common-sense reasoning and coreference resolution](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. 2023. [Llm blueprint: Enabling text-to-image generation with complex and detailed prompts](#).
- Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. [A framework for few-shot language model evaluation](#).
- Team AI Google. 2023. Bard. <https://blog.google/technology/ai/bard-google-ai-search-updates/>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#).
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.
- Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. [Mitigating gender bias amplification in distribution by posterior regularization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2936–2942, Online. Association for Computational Linguistics.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

- Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Stefan Eger. 2023. Towards explainable evaluation metrics for machine translation. *arXiv preprint arXiv:2306.13041*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Truthfulqa: Measuring how models mimic human falsehoods](#).
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a large annotated corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Larry R Medsker and LC Jain. 2001. Recurrent neural networks. *Design and Applications*, 5(64-67):2.
- Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.
- Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.
- OpenAI. 2023. [Gpt-4 technical report](#).
- John C Paolillo and Anupam Das. 2006. Evaluating language statistics: The ethnologue and beyond. *Contract report for UNESCO Institute for Statistics*.
- Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. [The lambada dataset: Word prediction requiring a broad discourse context](#).
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. [Language models as knowledge bases?](#)
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. [Photorealistic text-to-image diffusion models with deep language understanding](#).
- Andrea Santilli and Emanuele Rodolà. 2023. [Camoscio: an italian instruction-tuned llama](#).
- Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. 2020. [Super-glue: Learning feature matching with graph neural networks](#).
- Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.
- Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13693–13696.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023a. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023b. Stanford alpaca: An instruction-following llama model.

- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2021. From natural language processing to neural databases. In *Proceedings of the VLDB Endowment*, volume 14, pages 1033–1039. VLDB Endowment.
- Gabriele Tolomei, Cesare Campagnano, Fabrizio Silvestri, and Giovanni Trappolini. 2023. [Prompt-to-os \(p2os\): Revolutionizing operating systems and human-computer interaction with integrated ai generative models](#). In *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 128–134.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Rangan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).
- Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, Alon Halevy, and Fabrizio Silvestri. 2023. Multimodal neural databases. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2619–2628.
- Daniel Trippa, Cesare Campagnano, Maria Sofia Bucarelli, Gabriele Tolomei, and Fabrizio Silvestri. 2024. [\$\nabla\tau\$: Gradient-based and task-agnostic machine unlearning](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Eric Wallace, Matt Gardner, and Sameer Singh. 2020. Interpreting predictions of nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 20–23.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.
- Canwen Xu and Julian McAuley. 2023. [A survey on model compression and acceleration for pretrained language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10566–10575.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.