Teasing LLMs adapted to Italian

Leonardo Ranaldi[1,2], Giulia Pucci[1], Elena Sofia Ruzzetti[1], Fabio Massimo Zanzotto[1] and André Freitas[2,3]

[1]Università degli studi Roma Tor Vergata, Italy
[2]Idiap Research Institute, Switzerland
[3]Department of Computer Science, University of Manchester, UK

Abstract

Instruction-tuned Large Language Models (It-LLMs) are changing NLP thanks to their easy accessibility. These models seem able to grasp language, solve complex tasks, and perform even with few resources. These abilities and ease of handling democratize their use, enabling many researchers to produce their homemade It-LLMs. However, the complete understanding of their potential needs to be improved due to the black-box nature of many models and the absence of holistic evaluation studies. We present an evaluation resource for It-LLMs tuned in Italian to address these challenges. Our proposal includes evaluating models on several aspects. We take a holistic approach to analyzing model performance factors, including the pre-training base, instruction-tuning data, and training methods. Our results reveal that data quality is the most crucial factor in scaling model performance. While available open-source models demonstrate impressive ability, they present problems when customized adapters are used. We are encouraged by the rapid development of models by the open-source community. However, we also highlight the need for rigorous evaluation to support the claims.

Keywords

Instruction-tuned Large Language Models, Multilingual LLMs,

1. Introduction

The advent of Instruction-tuned Large Language Models (It-LLMs) marks yet another change in NLP in the last few decades. Indeed, their abilities are evident in numerous applications, from complex problem-solving to information retrieval to conversational assistants such as ChatGPT. Examples include GPT-4, which demonstrates abilities in language comprehension and common sense, logical-mathematical problem solving, law, and medicine. However, despite their remarkable competence and adaptability, the full extent of their potential has yet to be fully understood. Indeed, their direction is poorly captured, given many models' simple use, black-box nature, and lack of in-depth and holistic evaluation studies [1, 2, 3].

To manage these challenges and deeper understand the abilities of these models, a series of evaluation benchmarks were introduced that are explicitly designed for the comprehensive evaluation of It-LLMs [4, 5, 6, 7, 8, 9]. However, evaluation resources are only available in English, and it is tricky and misleading to evaluate a model trained on instructions in the Italian language.

In this paper, we propose evaluation resources for Italian It-LLMs. Furthermore, we tested a set of open-source It-LLMs fine-tuned in the Italian language, demonstrating excellent adaptability and some gaps in downstream performance. In particular, our methodology, applying a systematic and holistic approach, examines the problem-solving ability, writing ability, and alignment between languages of customized It-LLMs that are fine-tuned in a specific language, i.e., Italian, starting from the work proposed by Chia et al. [5]. Through a rigorous exploration of these factors, we seek to shed light on the vital elements that determine the performance of the models, facilitating an understanding of how these models can best be harnessed to meet our needs. Our contribution is fully available and open-source [1]

2. The Open-Source Instructed LLMs

Large Language Models (LLMs) have caught mainstream attention; they have become a comprehensive category of models. LLMs are comprehended as pre-trained and fine-tuned models with general language prompts or Instruction-tuned models. Therefore, we distinguish between basic and Instructed models, where basic LLMs are pre-trained LLMs that can be fine-tuned on instructions to become Instruction-tuned LLMs (It-LLMs). In particular, in Table 1, we summarize mainly open-source

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy
name.username@idiap.ch (L. Ranaldi);
name.username@uniroma2.it (G. Pucci);
name.username@uniroma2.it (E. S. Ruzzetti);
name.username@uniroma2.it (F. M. Zanzotto);
name.username@idiap.ch (A. Freitas)

[1] **https://github.com/LeonardRanaldi/italian-instruct-eval**

Model	*Architecture*	*num. Tokens*	*Source*
Llama	Decoder	1.4T	Unknown∗
Llama-2	Decoder	2.4T	Unknown∗
OPT	Decoder	180B	The Pile
BLOOM	Decoder	250B	Unknown∗
T5	Encoder-Decoder	1T	C4

Table 1
Open-source Large Language Models, with ∗ we denote data dump not available.

Model	**Backbone**	**Size**	**Source**	**Training**
Alpaca [11]	LLaMA	7-30B	Alpaca Data	Supervised
Baize [13]	LLaMA	7-30B	Self-Chat Data	Supervised
Vicuna [14]	LLaMA	7-33B	ShareGPT	Supervised
Falcon [15]	LLaMA	7-40B	Refinedweb	Supervised
ChatGLM [16]	GLM	6B	Unknown	RLHF
Customized Adapter				
Camoscio [17]	LLaMA	7B	Alpaca(Italian)	Supervised
Stambecco [18]	LLaMA	7-13B	Alpaca(Italian)	Supervised
Fauno [19]	LLaMA	7-13B	Baize data(italian)	Supervised

Table 2
Details of open-source instructed LLMs.

LLMs due to the need for more transparency and reproducibility of closed-source models.

The essential part of the Instruction-tuning idea is the data used to train LLMs. Indeed, factors such as quality, quantity, and format can determine the behavior of the instructed model. Table 3 presents several open-source resources. There is a growing tendency to exploit synthetic instruction data from closed-source models [10, 11]. While this practice can allow instructed models to mimic the behavior of closed-source models, it can also lead to problems such as the inheritance of the black-box nature of closed-source models and instability due to noisy synthetic instructions [12].

Finally, a holistic overview of the instructed open-source models can be found in Table 2, where the basic model with dimensions, instruction dataset, and training method for each It-LLMs is given. We observe a variety of model sizes and datasets. Therefore, this overview of open-source instructed LLMs provides comprehensive factors for evaluation and analysis.

3. Challenges & Methods in Evaluating Intruction-tuned LLMs

3.1. Background and Challenges

The highest wall in evaluating LLMs is the closed-source concept, where creators often hide model details, instruction datasets, and training methods. Such models thus lead to a knowledge vacuum in the research community as it is impossible to rigorously analyze the reasons for their performance.

On the other side of the coin is an ongoing open-source development that aims to democratize language model technology. While these efforts are highly encouraged, the pace of development of new models can outpace advances in evaluation studies. Unfortunately, informal evaluations often spot new models, which must be clarified when comparing different models.

We should consider different factors, such as pre-training and instruction data, to arrive at a holistic understanding of LLMs and It-LLMs. While previous work has conducted in-depth studies in some areas, such as datasets [20] and more concrete, such as general benchmarks [21], other factors should be considered to achieve a complete understanding. For example, performances on customized models for particular languages or tasks.

Recent work shows the elasticity and customization of It-LLMs in many languages. Santilli and Rodolà [17] translated Alpaca [11] into Italian by proposing Camoscio. Later, in Stambecco, the author [18] reproduced the same work by modifying some parameters. In [19], the models of the Baize [13] family were adapted into Italian. In this new scenario, evaluation has become increasingly important and challenging. Recent evaluation studies produce concrete results such as accuracy and precision [5, 22]. However, these methodologies are generic and not customized for a specific task and language.

Model	*Size*	*Domain*	*Source*
Alpaca Data [11]	52K	General	GPT-3
Self-Instruct [10]	52K	General	Human-Annotation
ShareGPT [14]	70K	Dialogue	ChatGPT
Self-Chat [13]	100k	Dialogue	ChatGPT

Table 3
Open-source Instruction-tuning datasets.

Finally, Ranaldi et al. [4], generalizing previous work, proposed a cross-lingual approach by eliciting It-LLMs with multilingual Alpaca empowered with translation-following demonstrations.

In this paper, we propose an Italian evaluation method for Italian fine-tuned It-LLMs. Our method is based on various general skills and usage scenarios applicable to It-LLMs adapted.

3.2. Proposed Methods

We propose to translate three well-known resources to evaluate the abilities of several Intruction-tuned Large Language Models. To perform well, the adapted models should have inherited world awareness, multi-hop reasoning, and more, merely like the original models. These benchmarks are:

Massive Multitask Language Understanding (MMLU) [23] measures knowledge of the world and problem-solving problems in multiple subjects with 57 subjects across STEM, humanities, social sciences, and other areas.

	Model	MMLU		BBH		DROP	
		Acc.	δ	Acc.	δ	Acc.	δ
Original data	∗Alpaca-Lora 7B	35.6	-	30.7	-	27.5	-
	♣Alpaca-Lora 13B	50.9	-	32.6	-	31.8	-
	+ Alpaca-Lora 30B	58.4	-	41.3	-	45.1	-
	∘Baize 7B	43.5	-	45.6	-	53.8	-
	⋄Baize 13B	50.9	-	49.5	-	56.4	-
	- Baize 30B	59.8	-	64.6	-	69.8	-
Italian data	∗Alpaca-Lora 7B	35.1	-0.5	30.1	-0.6	26.9	-0.6
	♣Alpaca-Lora 13B	50.6	-0.3	32.1	-0.5	31.6	-0.2
	+ Alpaca-Lora 30B	57.9	-0.5	41.1	-0.2	44.9	-0.2
	∘Baize 7B	44.3	-0.8	46.3	-0.7	54.5	-0.7
	⋄Baize 13B	51.2	-0.3	49.8	-0.6	57.2	-0.8
	- Baize 30B	59.5	-0.5	65.2	-0.4	70.1	-0.3
	Italian Adapters 7B	–	–	–		–	–
	∗Camoscio 7B	30.2	**-5.4**	29.8	*-0.8*	22.0	**-4.5**
	∗Stambecco 7B	28.2	**-7.4**	29.7	*-0.9*	21.6	**-5.9**

Table 4
Evaluation results. We denote the accuracy across the benchmarks as Acc., while δ denotes the performance change compared to the original version trained and evaluated on English datasets.

Discrete Reasoning Over Paragraphs (DROP) [24] reading comprehension on mathematics where the model should perform discrete reasoning on passages extracted from Wikipedia articles.

BIG-Bench Hard (BBH) [25] is a subset of challenging tasks related to navigation, logical deduction, and fallacy detection.

Evaluation Each benchmark was translated into Italian using google API [2]. Then, zero-shot evaluations were done for the original version in English and ours in Italian using the framework proposed in [5] [3].

4. Results

Customized Instruction-tuned Large Language Models (It-LLMs) need further refinement. This statement is supported by the results shown in Table 4 on fine-tuned models over the Italian benchmarks. Firstly, the original Alpaca-Lora and Baize evaluated on the English benchmark outperformed Camoscio, Stambecco, and Fauno evaluated on the Italian benchmark.

Secondly, the differences between Camoscio, Stambecco, Fauno, and the original Alpaca-Lora and Baize are very close on the Italian benchmarks (Italian Data on Table 4). Thirdly, models with more parameters (30B) performed best, and the δ between performances on English-language benchmarks are remarkably lower than models with fewer parameters. In conclusion, fine-tuning a customized resource, in this case, customized English language resources, was insufficient to increase performance. This phenomenon may be due to the quality of the data used for homemade fine-tuning and also suggests that fine-tuning on custom It-LLMs may have inserted a bias. These gaps should be further investigated, and the scientific community should pay more attention.

5. Conclusions

In this paper, we have presented a systematic evaluation of four resources for Instruction-tuned Large Language Models (It-LLMs). Our holistic approach analyzed critical performance factors and showed that efforts to customize It-LLMs are not always rewarded by performance.

Underlining the importance of the contribution of the open-source community in proposing new solutions to meet specific needs. We emphasize the significance of data quality in scaling model performance. Additionally, our translated benchmarks provide valuable insights into the adaptability and effectiveness of It-LLMs for specific language tasks. By addressing key evaluation challenges, our work contributes to the responsible and effective utilization of It-LLMs, fostering further advancements in NLP.

In future developments, we will investigate lightweight approaches to elicit adapters' multi- and cross-lingual skills inspired by what has been done in [4, 9]

[2] available here https://github.com/LeonardRanaldi/italian-instruct-eval

[3] https://github.com/declare-lab/instruct-eval

References

[1] L. Ranaldi, E. S. Ruzzetti, F. M. Zanzotto, Precog: Exploring the relation between memorization and performance in pre-trained language models, 2023. arXiv:2305.04673.

[2] L. Ranaldi, E. S. Ruzzetti, D. Venditti, D. Onorati, F. M. Zanzotto, A trip towards fairness: Bias and de-biasing in large language models, 2023. arXiv:2305.13862.

[3] L. Ranaldi, A. Nourbakhsh, A. Patrizi, E. S. Ruzzetti, D. Onorati, F. Fallucchi, F. M. Zanzotto, The dark side of the language: Pre-trained transformers in the darknet, 2022. arXiv:2201.05613.

[4] L. Ranaldi, G. Pucci, A. Freitas, Empowering cross-lingual abilities of instruction-tuned large language models by translation-following demonstrations, 2023. arXiv:2308.14186.

[5] Y. K. Chia, P. Hong, L. Bing, S. Poria, Instructeval: Towards holistic evaluation of instruction-tuned large language models, 2023. arXiv:2306.04757.

[6] L. Ranaldi, G. Pucci, F. M. Zanzotto, Modeling easiness for training transformers with curriculum learning, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 937–948. URL: https://aclanthology.org/2023.ranlp-1.101.

[7] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, 2023. arXiv:2307.03109.

[8] L. Ranaldi, G. Pucci, Knowing knowledge: Epistemological study of knowledge in transformers, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/2/677. doi:10.3390/app13020677.

[9] L. Ranaldi, F. M. Zanzotto, Empowering multi-step reasoning across languages via tree-of-thoughts, 2023. arXiv:2311.08097.

[10] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, H. Hajishirzi, Self-instruct: Aligning language models with self-generated instructions, 2023. arXiv:2212.10560.

[11] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[12] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, W. Chen, Synthetic prompting: Generating chain-of-thought demonstrations for large language models, arXiv preprint arXiv:2302.00618 (2023).

[13] C. Xu, D. Guo, N. Duan, J. McAuley, Baize: An open-source chat model with parameter-efficient tuning on self-chat data, arXiv preprint arXiv:2304.01196 (2023).

[14] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. URL: https://lmsys.org/blog/2023-03-30-vicuna/.

[15] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, G. Penedo, Falcon-40B: an open large language model with state-of-the-art performance (2023).

[16] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, J. Tang, Glm-130b: An open bilingual pre-trained model, 2022. arXiv:2210.02414.

[17] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, 2023. arXiv:2307.16456.

[18] Michael, Stambecco: Italian instruction-following llama model, https://github.com/mchl-labs/stambecco, 2023.

[19] A. Bacciu, G. Trappolini, A. Santilli, E. Rodolà, F. Silvestri, Fauno: The italian large language model that will leave you senza parole!, 2023. arXiv:2306.14457.

[20] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, A. Roberts, The flan collection: Designing data and methods for effective instruction tuning, 2023. arXiv:2301.13688.

[21] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, N. Duan, Agieval: A human-centric benchmark for evaluating foundation models, 2023. arXiv:2304.06364.

[22] J. Sun, C. Shaib, B. C. Wallace, Evaluating the zero-shot robustness of instruction-tuned language models, 2023. arXiv:2306.11270.

[23] D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt, Measuring mathematical problem solving with the math dataset, 2021. arXiv:2103.03874.

[24] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, M. Gardner, DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 2368–2378. URL: https://aclanthology.org/N19-1246. doi:10.18653

/v1/N19-1246.

[25] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, J. Wei, Challenging big-bench tasks and whether chain-of-thought can solve them, 2022. arXiv:2210.09261.