PhD proposal #2
2024

Evaluation and Assessment of Reasoning in Large Language Models & Foundation Models

AXA-Sorbonne University Trustworthy and Responsible AI Lab (TRAIL)

Supervision:

Keywords: LLM, Large Language Model, Foundation Model, Reasoning, Evaluation, Assessment, Trustworthy and Responsible Machine Learning

More information & Application: `trail@listes.lip6.fr` - `https://trail.lip6.fr`

1 Context

Reasoning is the process of deriving conclusions from evidence and logic. Reasoning stands as a distinctive characteristic associated with human intelligence. This cognitive ability is critical to address complex tasks as it is used for solving problems, making decisions and planning. Multiple scientific disciplines such as philosophy, psychology, mathematics, and computer science have contributed to understand human reasoning and to develop techniques to model reasoning and to replicate its mechanisms artificially.

Large Language Models (LLMs) and Foundation Models (FMs) [1, 2, 15, 7] mark a significant step to reproduce, or mimic, reasoning. LLMs are trained on vast corpus of data, leveraging technical advances in machine learning such as the attention mechanism and the transformer [17]. The emergence of reasoning abilities is among their most significant breakthroughs, [14, 4, 13]. These abilities are thought to emerge as model size increases, particularly beyond 100 billion parameters [18].

This emergence is intriguing as these abilities appear without being explicitly brought into the LLMs, through an objective function or explicit programming [8]. Also, reasoning abilities coexist with LLMs' performances in language understanding and generation, making them more versatile, capable to address more complex tasks and in a user-friendly way [6]. Reasoning abilities in LLMs allow to envision, machine learning models capable to address human-complex tasks.

However, the exact nature of LLMs' reasoning abilities is still an open question: is it a mimicry of human reasoning or an authentic form of reasoning? [10] To what extent can we trust LLMs' reasoning abilities? Despite their successes, LLMs still stumble across complex reasoning tasks [16, 11]. As these models could be leveraged in many applications requiring reasoning, it is imperative to deepen our understanding of their reasoning capabilities, how LLMs actually address reasoning, what are the limits and how LLMs' reasoning abilities can be improved through methods like prompting [19] and fine-tuning [3, 20, 5] and critically how we can accurately assess these abilities [16]. These questions apply in particular to new or particular contexts and use-cases.

2 Scientific Objectives

During this PhD, the main objective will be to address the question of the assessment and the evaluation of the reasoning abilities of LLMs.

Many benchmarks have been developed to assess NLP (Natural Language Processing) and NLU (Natural Language Understanding) tasks, which have been gathered under meta-benchmarks to assemble many tasks to assess the versatile abilities of LLMs [12, 9], in particular to evaluate different views on reasoning. We are interested in particular by the question of the assessment

and the evaluation of the reasoning abilities of LLMs applied to novel use-cases where classical benchmarks cannot be blindly trusted for an adequate assessment.

Following a literature review on the multi-facets of reasoning, reasoning tasks and how these are addressed and assessed in Language Models and LLMs, the objective will be to define the reasoning abilities required to solve specific use-cases. That may involve the study of the human expectations for reasoning in LLMs (users and developers). With the identification of gaps in the state of the art, novel techniques will be developed to better assess and evaluate LLMs' reasoning abilities. These techniques will have to be themselves critically assessed.

Challenging use-cases involving reasoning tasks will guide the study, such as the automation of contract coverage check for insurance customers with LLMs.

Following the development of adequate assessments of the reasoning abilities, a study of LLMs reasoning limits will be performed. This research will also explore the link with interpretability, aiming to delve deeper into understanding LLMs reasoning, with contributions on the reliability and transparency of these models. Finally, this work could lead to propositions to improve LLMs' reasoning abilities (in particular on a specific use-case).

3 Expected Contributions

During the thesis, the PhD candidate is expected to produce research articles to be submitted to high-quality peer-reviewed ML workshops, conferences and journals (e.g. ICML, IJCAI, NeurIPS, JMLR...). Algorithmic implementations of the conceived methodology will be made available through open-source libraries.

4 Working Environment

This PhD is hosted by the joint research lab TRAIL between AXA and Sorbonne University, in Paris. As such, the PhD Candidate will be hired by Sorbonne University and supervised by TRAIL members from Sorbonne University and the AI research team from AXA in Paris.

The PhD Candidate will also benefit from interactions with other researchers from the TRAIL ecosystem, gathering research expertise in NLP, deep learning, Responsible AI and Human-Computer Interactions. Besides the Sorbonne University campus, other researchers from TRAIL are based on the EPFL campus in Lausanne (Switzerland) and the Stanford University campus in Palo Alto (US). Depending of the thesis advancement, a collaboration and potential short research stay in these universities may be considered.

5 Profile and skills

- MSc. in computer science, applied mathematics or equivalent, with in-depth coverage of the artificial intelligence and machine learning fields.
- Study and practical experience in implementing and using language models, large language models on NLP and NLU tasks.
- Interest for the linguistics and cognitive sciences fields.
- Previous experience in research: research project, internships, etc.
- Good experience in programming in python and ML libraries (especially for NLP, LMs and LLMs).
- Advanced level in English.

References

[1] 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*, 33:1877–1901, 2020.

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[3] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[4] J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, 2023.

[5] J. Huang, S. S. Gu, L. Hou, Y. Wu, X. Wang, H. Yu, and J. Han. Large language models can self-improve. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[6] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022.

[7] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

[9] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[10] A. Patel, S. Bhattamishra, and N. Goyal. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, 2021.

[11] L. Ruis, A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*, 2022.

[12] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[13] J. Sun, C. Zheng, E. Xie, Z. Liu, R. Chu, J. Qiu, J. Xu, M. Ding, H. Li, M. Geng, et al. A survey of reasoning with foundation models. *arXiv preprint arXiv:2312.11562*, 2023.

[14] M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. Le, E. Chi, D. Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

[15] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[16] K. Valmeekam, A. Olmo, S. Sreedharan, and S. Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[18] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[20] E. Zelikman, Y. Wu, J. Mu, and N. Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.