PhD proposal #1
2024

Trustworthy Large Language Models with Natural Language Instructions
AXA-Sorbonne University Trustworthy and Responsible AI Lab (TRAIL)

Supervision:

Keywords: LLM, large language model, trustworthy and responsible machine learning

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

1 Context

Large Language Models (LLMs) Recently, large language models (LLMs) have gained popularity due to their ability to generate coherent and contextually appropriate text. Building on the recent Transformer architecture [18], LLMs have thus rapidly been identified as opening new opportunities, largely fostered by their novel generation capabilities along with their versatility [5, 19]. This popularity has led to a soaring number of diverse AI applications being identified, including in critical tasks such as healthcare and finance. Yet, these opportunities also come with the downside of being extremely large, and therefore difficult to control, which as a result has led to LLMs being pointed out as bearing new risks [5, 7, 1]: opaqueness, bias amplification, robustness, etc.

Building trustworthy LLMs As an answer to these questions, the field of Responsible AI focuses on topics such as ML Explainability (understanding ML systems outputs and behaviors), Algorithmic Fairness (measuring and mitigating bias and discrimination in AI models) and robustness (ensuring good capabilities in unseen environments). In particular, the concept of AI Alignment [9] addresses the pivotal challenge of ensuring that AI technologies, in particular Foundation Models, not only perform effectively but also operate in harmony with our collective values [12], societal principles [4], and the nuanced expertise inherent in various domains [10, 17]. Numerous contributions have been made in this direction recently, with most of them relying on fine-tuning models using vast amount of data [24, 8, 4] or in-context learning [6]. However, further studies have shown that they still fail to solve the issue completely [3, 22, 14]. Besides, they present the downside of being somewhat impractical, as relying on vast amount of raw feedback is costly, uncontrollable and untractable, therefore leaving the possibility of being prone to issues such as bias, lack of robustness, etc.

2 Scientific Objectives

This PhD proposal aims to further delve into trustworthy AI alignment, exploring novel methodologies to bridge the gap between the capabilities of LLMs and ethical considerations, along with expert knowledge, essential for their responsible deployment. In particular, this research seeks to see how models can be aligned through a novel type of input, in the form of task-related constraints formulated in natural language by a domain expert, thus expanding the range of alignment possibilities. Beyond traditional applications of AI alignment, some examples of use-cases where this methodology would be particularly helpful are:

- Bridge the gap between algorithmic and ethical notions of fairness by having an LLM comply to the fairness definition defined by an ethicist or lawyer.
- Enrich text classification methods by constraining input-output relationships that may not be necessarily reflected in the data. For instance, force causal connections between some concepts.

- Have a LLM answer questions using specific rules to answer. For instance, in a legal context, the IRAC framework [1] is taught in Law School to guide answers to legal questions. This could be used to enrich a LLM with legal knowledge.

Formulated in natural language by a domain expert, instructions would thus allow to convey much stronger and richer contextual information than what would have been easily available by relying solely on labelling of new data points, as proposed by existing works. Leveraging these instructions, the objective of the Ph.D. is then to develop a methodology to guide the model's learning and decision process. Ultimately, the goal of this thesis is to contribute to the development of robust and ethically aligned AI technologies, broadening the range of alignment possibilities.

Some key problems to address include:

- **Making instructions computable.** Leveraging natural language instructions in the training raises the question of the form in which they should be integrated in the learning and decision process of the model. How to make these instructions computable? While injecting the instructions as raw text remains an option, other areas that may be interesting to explore for this purpose include (but are not limited to): concept learning [13, 21, 2], which allows the computation of user-understandable concepts (e.g. "stripes" to define a zebra) as learnable features. While a few attempts have been made to adapt this notion to NLP [**?**], these works remain preliminary ; code generation from text intructions (e.g. [15]), that would allow the translation of textual instructions into executable modules for increased control and transparency.

- **Integrating and evaluating instructions.** Once instructions are made computable, the questions of how to use this knowledge effectively is to be explored. Although the exact nature of this integration shall be defined in adequacy with the solution considered in the previous task, some areas to explore include e.g. adversarial debiasing [23] and output self-correction [20]. Another promising direction to investigate is model editing (see e.g. [11]) and mechanistic interpretability [16], which aim at interpreting and controlling model behaviors by interacting directly with their neuronal components. code generation from text intructions (e.g. [15]), that would allow the translation of textual instructions into executable modules for increased control and transparency.

- **Instruction Definition:** The very definition of the expert instructions obviously represents a crucial question. Leveraging the knowledge acquired while studying the first two problems, the objective hereby described is to understand how to properly formulate, and collect, user feedback. Questions such as overlap between instructions and language ambiguity may also be investigated.

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. Finally, application of the methodology to a real-world usecase may be envisaged depending on the findings and the profile of the candidate.

[1] https://www.iracmethod.com/irac-methodology

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 Université 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 required

Required:

- MSc. in computer science, AI, data science, applied mathematics, statistics or equivalent.
- Good experience in programming in python and ML libraries (e.g. pytorch, scikit-learn...)
- Good knowledge of deep learning, machine learning, statistical modelling, or equivalent
- Advanced level in English (technical discussions, presentations and paper writing are expected)

Preferred:

- Previous research experience: research projects, internships, publications...
- Experience with pytorch, scikit-learn, Hugging Face...
- Experience in NLP, especially LLMs

References

[1] C. Barrett, B. Boyd, E. Burzstein, N. Carlini, B. Chen, J. Choi, A. R. Chowdhury, M. Christodorescu, A. Datta, S. Feizi, et al. Identifying and mitigating the security risks of generative ai. *arXiv preprint arXiv:2308.14840*, 2023.

[2] K. Bhatia, A. Narayan, C. M. De Sa, and C. Ré. Tart: A plug-and-play transformer module for task-agnostic reasoning. *Advances in Neural Information Processing Systems*, 36:9751–9788, 2023.

[3] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023.

[4] F. Bianchi, M. Suzgun, G. Attanasio, P. Röttger, D. Jurafsky, T. Hashimoto, and J. Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. *arXiv preprint arXiv:2309.07875*, 2023.

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] 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.

[7] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, Z. He, Y. Duan, M. Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 651–666, 2023.

[8] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[9] I. Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[10] N. Guha, J. Nyarko, D. E. Ho, C. Re, A. Chilton, A. Narayana, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[11] P. Hase, M. Bansal, B. Kim, and A. Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

[13] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[14] F. Ladhak, E. Durmus, M. Suzgun, T. Zhang, D. Jurafsky, K. Mckeown, and T. B. Hashimoto. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3198–3211, 2023.

[15] Y. Li, D. Choi, J. Chung, N. Kushman, J. Schrittwieser, R. Leblond, T. Eccles, J. Keeling, F. Gimeno, A. Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[16] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[17] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[18] 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.

[19] 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.

[20] T.-H. Wu, L. Lian, J. E. Gonzalez, B. Li, and T. Darrell. Self-correcting llm-controlled diffusion models. *arXiv preprint arXiv:2311.16090*, 2023.

[21] M. Yuksekgonul, M. Wang, and J. Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2022.

[22] Q. Zhan, R. Fang, R. Bindu, A. Gupta, T. Hashimoto, and D. Kang. Removing rlhf protections in gpt-4 via fine-tuning. *arXiv preprint arXiv:2311.05553*, 2023.

[23] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

[24] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.