A Big Challenge: Tools to Guarantee Robust and Controlled Behavior of Large Language Models

Fabio G. Cozman, Sarajane M. Peres, Marcelo Finger, Renata Wassermann, Anna H. Reali Costa, Edson S. Gomi, Artur J. L. Correia, Anarosa A. F. Brandão, Karina V. Delgado, Denis D. Mauá, Thiago A. S. Pardo, Fátima L. S. N. Marques

[1]Universidade de São Paulo (USP)
Av. Profº Lúcio Martins Rodrigues, 370 – 05508-020 — São Paulo — SP – Brazil

{fgcozman,sarajane,anna.reali,gomi,arturjordao,kvd}@usp.br

{anarosa.brandao,denis.maua,fatima.nunes}@usp.br

{mfinger,renata}@ime.usp.br, taspardo@icmc.usp.br

Abstract. *Large language models (LLMs) have changed Artificial Intelligence, and in fact greatly affected Computer Science and its applications. Even though the capabilities of LLMs are impressive, they offer responses without any solid guarantees of rationality, are prone to hallucinations, are relatively weak when faced with long reasoning chains, and offer a limited degree of controllability. Despite the impressive performance, ensuring robust and controlled output is a major challenge. That is, the big challenge is to produce tools that make these models insensitive to irrelevant features, resistant to unexpected failures, amenable to control and in accordance to performance requirements, as well as tools to formally verify their success and failure modes, and to evaluate them in meaningful ways.*

Resumo. *Grandes modelos de linguagem transformaram a área da Inteligência Artificial e, de fato, impactaram profundamente a Ciência da Computação e suas aplicações. Embora suas capacidades sejam notáveis, esses modelos produzem respostas sem garantias sólidas de racionalidade, são suscetíveis a alucinações, demonstram fragilidade diante de cadeias longas de raciocínio e oferecem um grau limitado de controlabilidade. Apesar de seu desempenho impressionante, assegurar que suas saídas sejam robustas e controladas continua sendo um desafio. Nesse contexto, portanto, o grande desafio é desenvolver ferramentas que tornem esses modelos insensíveis a características irrelevantes, resistente a falhas inesperadas, passíveis de controle e conforme com requisitos de desempenho, além de instrumentos capazes de verificar formalmente seus modos de acerto e falha e de avaliá-los de maneiras que façam sentido.*

1. Introduction

Large language models (LLMs)[1] have had extraordinary impact. They were developed at first to convey statistical properties of languages [Jurafsky and Martin 2024], but re-

[1]For a curated and updated list of recent work, see https://github.com/Hannibal046/Awesome-LLM

cent developments have demonstrated that LLMs can mimic human conversation to astonishing levels, leading some to think that they offer a direct path to Artificial General Intelligence [Bubeck et al. 2023].

Alas, whatever mechanisms are at play within LLMs, the research community can hardly say they are fully understood. For instance, the community does not know when LLMs will fail, and how to make them withstand changes; it knows LLMs hallucinate, but does not know exactly when they do it, how they do it, and how this behavior can be blocked or at least how it can be guaranteed to stay within given required bounds. Moreover, while LLMs often succeed in seemingly human ways, for example by correctly summarizing documents, they fail in surprising non-human ways, for example by missing easy questions while nailing hard questions in a given exam [Locatelli et al. 2024].

At this point the research community does not have the right tools to analyze and synthesize them so as to guarantee they satisfy a set of requirements; does not know how to guarantee that their responses follow given rules; and does not know how to make them reason in formally guaranteed ways, for instance by staying within prescribed logical schemes. Moreover, there is little understanding on how to design LLMs, other than by repeating a limited set of architectures that have been successful in the past.

This context presents us with a number of challenges, ranging from technical to ethical issues. However, we wish here to focus on one key challenge that is directly related to the theme of this meeting: ***the development of theoretical and practical computational tools that can ensure, or at least significantly improve, the robustness and control of LLMs' behaviour.***

2. LLMs with robust and controlled behaviour

First, there is a need for basic research on the mathematical tools to understand the inner workings of LLMs. To be able to control something, it is important to understand it; hence there must be a better grasp of the relationship between complexity and expressivity, of optimization algorithms and performance, of architecture size/structure and robustness to failure. Besides, it is necessary to determine some paradigmatic problems in this effort, much as resolution for first-order logic or completeness for automata offer a guiding path to investigation. These aspects depend on connections with research in statistics, statistical physics, and mathematics; similar multi-disciplinary efforts have been pursued throughout the world (for instance, take the recent call for proposals by the US-NSF.[2])

There is also a need for a systematic study of LLM architectures that guarantee correctness and assertiveness of outputs, both in the context of their fundamental task of language generation (in an intrinsic perspective) and in well-established contexts of downstream applications (in an extrinsic perspective) [Bommasani et al. 2021]. These architectures must be studied with respect to metrics such as accuracy, but also with respect to their robustness and reliability.

One particular strategy to enhance LLMs has been Retrieval-Augmented Generation (RAG) [Gao et al. 2024], where a LLM may query external databases. This sort of strategy has been expanded to include queries to reasoning engines, often supported

[2]Program Solicitation NSF 24-569, https://new.nsf.gov/funding/opportunities/mfai-mathematical-foundations-artificial-intelligence

by complex prompting techniques that ask for the LLM to expose its chain of thought — indicating steps that may require external reasoners. On one hand, reliance on formal reasoning is a promising approach to minimizing hallucinations and maximizing control through explicit rules. On the other hand, beyond the complexity of designing such formal reasoning, efforts to enhance robustness and control inevitably introduce trade-offs — particularly in terms of flexibility, creativity, and computational efficiency. Stricter control mechanisms may constrain model expressiveness or increase inference latency, potentially undermining usability in dynamic, real-time applications. Quantifying and balancing these dimensions remains a significant and specific research challenge that depends heavily on the intended application domain. Alas, current RAG systems are still far from guaranteeing specific levels of performance.

Actually, RAG offers one possible strategy within *neuro-symbolic* approaches, where the goal is to mix the data-centric power of neurally inspired architectures with the knowledge-centric power of formal symbolic reasoning [Garcez and Lamb 2023]. Several different paths are possible here [Lamb et al. 2020]. One path consists of systems where a neural module calls a symbolic engine (RAG and its variants fit here). Another path explores the reverse idea: a symbolic engine that calls a neural module (say, an LLM) and processes its output. Yet another strategy is one where symbolic rules and formal constraints are used to help build a neural module so as to guarantee given requirements; a related scheme embeds rules and deductive patterns into numerical spaces, so as to enforce them in the same spaces where embeddings operate. All such approaches, and their possible combinations, deserve more study as their potential is still unfilled.

Another challenge to solve to make LLMs more applicable is to reduce their size and their complexity, while pursuing robust and controlled behavior. Current LLMs are very large and energy-hungry at both learning and inference time. To what extent is so much flexibility needed? A related question is whether LLMs could be more modular and hence easier to analyze and to design. The recent work on distillation, pruning, quantization and similar techniques to reduce the size of LLMs, make the resulting models even more opaque and do not add any formal guarantees or modularity. Note that the challenge posed herein goes beyond mere miniaturization. The goal is to achieve reductions in size and complexity while embedding guarantees — whether statistical, logical, or otherwise — that render smaller models as trustworthy and predictable as their larger counterparts.

Yet another key element in this challenge is the verification of LLMs. While most computing systems can today be verified by ever more powerful formalisms, there are few formal ways to verify neurally inspired systems, and they do not generally scale up [Preto and Finger 2023]. There is a sore need for guaranteed LLM verification by appropriate algorithms.

The concretization of the theoretical directions outlined above is essential and constitutes a fundamental part of the proposed challenge. Their technical feasibility hinges on advancements in model instrumentation, modular training pipelines, and access to explainability layers. Practical implementation may also depend on the development of middleware capable of enforcing symbolic constraints or verifying runtime behavior through formal methods — or through systematic and robust heuristic procedures.

3. Technological Autonomy and Local Restrictions

Brazil has already seen cases of LLM use in public sector conversational agents, legal document summarization, and policy analysis. These examples reinforce the feasibility and urgency of developing locally designed, controlled, and trusted solutions that provide performance guarantees. Furthermore, strengthening domestic R&D fosters workforce development and reduces dependence on foreign platforms with opaque mechanisms.

In this context, the development of robust and controllable LLMs is especially critical for Brazil, as the country's digital infrastructure, regulatory frameworks, and sociocultural conditions may differ substantially from those of highly digitized nations. Off-the-shelf foreign models often fail to meet local demands – whether due to language limitations, legal incompatibilities, or a lack of contextual alignment. To be effective, LLMs in Brazil must be adapted to local legislation (e.g., the LGPD), cultural specificities, and public service needs. Thus, any formalism for the verification, validation, and monitoring of LLMs should be effective both as a general framework and in local applications.

4. Evaluation

Of course, a challenge only makes sense when it is possible to determine whether it is met. There are two possible standards to which respect success can be evaluated here. First, our challenge will be met at a theoretical level when it becomes possible to analyze and synthesize LLMs against prescribed requirements. Second, our challenge will be met at an empirical level when concrete LLMs are able to reach prescribed performance on benchmarks and meet the usage requirements determined by the technical, organizational, and social contexts in which they will be applied.

Some requirements are easily expressed: we might ask a system never to return different answers if a question is formulated in distinct natural languages. And some metrics are obvious: the probability of returning the correct answer is important in a system that answers questions — there are, in fact, dozens of metrics that apply to natural language processing in general and to LLMs in particular [Liang et al. 2023]. However, we submit that finding precise formalisms so as to express requirements and metrics is *part of the challenge*, given the lack of guidance concerning requirements, and the need for more nuanced metrics that really capture semantics. It is necessary to compare existing formalims and metrics; agreement on how to evaluate LLMs is itself a tangible victory.

It is also important to have some agreed-upon testing scenarios. For instance, we believe an interesting scenario of extraordinary social impact is the *generation of correct arguments and the detection of false arguments* in public discourse. Arguments are complex objects that can be formally analyzed and validated. Other concrete applications include educational tutoring systems capable of verifying mathematical reasoning, legal assistants that must comply with jurisdiction-specific constraints, and healthcare conversational agents required to operate within ethical and regulatory boundaries. These scenarios offer fertile ground for controlled evaluations and the specification of formal requirements. Responding to these demands with guaranteed levels of performance is an important activity that can test the robustness and control of LLMs. We expect that a plethora of new scenarios will be developed in the next years.

Finally, understanding and verification within the theoretical or experimental scope will not be sufficient to ensure that society adequately perceives and benefits from

the scientific progress eventually achieved. In this sense, real-world and holistic evaluations will serve as a definitive measure to determine to what extent the proposed challenge and the solutions presented are necessary and sufficient to position LLMs as beneficial tools for society.

5. Conclusion

This paper sets a challenge focused on the development of theoretical and practical computing tools that *guarantee* levels of robustness and control for Large Language Models (LLM). We have commented on a number of specific research directions that may lead us to meet this challenge; it is possible that they also lead to other unanticipated strategies. But the goal is clear: the research community must develop novel LLMs that can satisfy formal requirements and follow formal rules with guarantees, perhaps of statistical nature. Once such guaranteed behavior is available, we can collectively find ways to use LLMs responsibly and to impose ethical rules. Also, closer work with industry can also speed up progress toward the proposed challenge. Industry partners often operate under stringent constraints, such as latency, compliance, cost, and risk minimization. That calls for robust control requirements, realistic testbeds, and for system-level monitors that can evaluate performance both during development and during real-world deployment scenarios.

Acknowledgements

F. G. Cozman is partially supported by CNPq grant Pq 305753/2022-3. M. Finger is partially supported by CNPq grant Pq 302963/2022-7. A. H. R. Costa is partially supported by CNPq grant Pq 312360/23-1. D. D. Mauá is partially supported by CNPq grant Pq 305136/2022-4 and FAPESP grant 2022/02937-9. The authors would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP) grant 2019/07665-4) and from the IBM Corporation.

Referências

Bommasani, R. et al. (2021). On the opportunities and risks of foundation models. *ArXiv*.

Bubeck, S. et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4.

Gao, Y. et al. (2024). Retrieval-augmented generation for large language models: A survey.

Garcez, A. d. and Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406.

Jurafsky, D. and Martin, J. H. (2024). *Speech and Language Processing*.

Lamb, L. C. et al. (2020). Graph neural networks meet neural-symbolic computing: A survey and perspective. In *Int. Joint Conf. on Artificial Intelligence*.

Liang, P. et al. (2023). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Locatelli, M. S. et al. (2024). Examining the behavior of LLM architectures within the framework of standardized national exams in Brazil. *arXiv preprint arXiv:2408.05035*.

Preto, S. and Finger, M. (2023). Proving properties of binary classification neural networks via łukasiewicz logic. *Log. J. IGPL*, 31(5):805–821.