**Prove me wrong: An LLM-based approach to linguistic theory verification**

Foundation models are self-supervised broad-domain neural networks trained on large datasets that can be adapted to various downstream tasks (Bommasani, Hudson, Adeli et al. 2022). For example, numerous generative language models, such as GPT-4 (Achiam, Adler, Agarwal et al. 2023), Phi-3 (Abdin, Jacobs, Awan et al. 2024), or Gemma (Mesnard, Hardin, Dadashi et al. 2024), review and generate code, handle task-oriented instructions and capable of knowledge representation and reasoning. While foundation models for natural language processing, or large language models (LLMs), are widely spread in applied linguistics, their usage in theoretical science is questionable, since they fail to infer complex linguistic information, such as copredication, compound nominals, cognitive content, and prepositional phrase attachment (Saba 2023). Nevertheless, the linguistic probing procedure shows that LLM representations, or embeddings, capture syntactic and semantic relationships (Hewitt, Manning 2019; Manning, Clark, Hewitt et al. 2020; Starace, Papakostas, Choenni et al. 2023).

Different approaches to applying LLMs to theoretical linguistics give contradictory results. LLM inference described in (Saba 2023) adapts the general-purpose text generation model to downstream tasks inspired by theoretical linguistics through prompt engineering. Since LLMs are not trained for specific linguistic tasks, they likely do not encode language interrelationships required to solve such tasks as linguistic structure parsing, resulting in poor performance in linguistic pattern inference.

The probing approach described in (Hewitt, Manning 2019) allows for programmatic extraction of linguistic interrelationships encoded in the LLM weights. This method accesses implicit connections through model embeddings and converts them to a structure, such as a graph representing a syntax tree. For example, calculating Euclidean distance between LLM word representations allows for linear transformation and mapping syntactic structure that can be compared to existing linguistic theory.

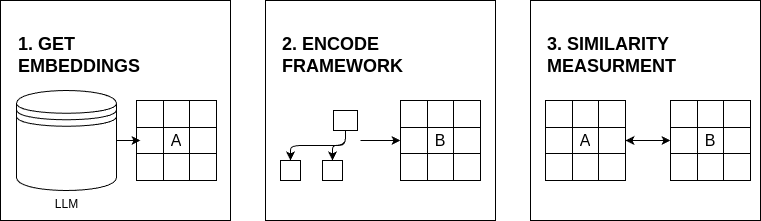
The latter approach proves that LLM embeddings often match Universal Dependencies categories (Nivre, de Marneffe, Ginter et al. 2020), meaning that neural networks likely encode linguistic structures similarly to linguistic theoretical grammar frameworks. However, it is not clear whether the structures restored from neural network weights would match other grammar frameworks, such as Head-Driven Phrase Structure Grammar (Pollard, Sag 1994), Construction Grammar (Fillmore, Kay 1999) or Lexical Functional Grammar (Bresnan 2001).

Considering the fact that LLM embeddings are the result of probabilistic calculations over real-life language data, it is possible to assume that a representational linguistic theoretical framework should correspond to the structure encoded in the neural network weights, such as the structure extracted through the probing procedure. The study hypothesis is that LLM word embeddings Euclidean space should match vector space based on theoretical grammar framework, if this framework is representative and valid. The study contributes to the field of theoretical linguistics providing a practical tool for linguistic theory development and verification.

The research method is the following. The first stage is to get embeddings from quantized LLM through llama.cpp library (Gerganov 2024) using an open-domain language dataset D. The next stage is to encode the theoretical grammar framework. The llama.cpp library uses Backus-Naur form (Backus 1959) as a formal grammar notation, however, the study proposes applying linear algebra instead. The study aims to develop a universal tool for linguistic theory verification, however, each grammar framework uses different notation. The study proposes using LLMs to synthesize linear equations from textual grammar descriptions and build vector space representation for the grammar framework on the dataset D. The final stage is Euclidean space mapping, which is a comparison of vector values between two spaces aiming to find exact matches, plausible similarities and dissimilarities between spaces.

Overall, the study proposes a novel research dataset for theoretical linguistics, a theory verification algorithm, and a universal formal grammar notation proposal. Figure 1 illustrates the proposed algorithm. The research codebase and the dataset is provided in the project repository (the link will be provided after the paper deanonymization).

Fig. 1. The proposed algorithm illustration

References

Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H. and Benhaim, A. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

Backus, J.W. The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM Conference. Proceedings of the International Conference on Information Processing. 1959. UNESCO. P. 125–132.

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

Bresnan, J. Lexical-functional syntax. Oxford: Blackwell, 2001.

Gerganov G. LLM inference in C/C++. URL: https://github.com/ggerganov/llama.cpp. Accessed 30.06.2024

Hewitt J., Manning C.D. A structural probe for finding syntax in word representations. 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), 2019. Minneapolis, Minnesota: Association for Computational Linguistics. P. 4129–4138.

Kay P., Fillmore C.J. Grammatical Constructions and Linguistic Generalizations: The What's X Doing Y? Construction. Language. Linguistic Society of America. 1999. 75 (1). P. 1–33.

Manning C.D., Clark K., Hewitt J., Khandelwal U., Levy O. Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences. 2020. 117(48). P. 30046–30054.

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J. and Tafti, P.. Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295, 2024.

Nivre J., de Marneffe M.-C., Ginter F., Hajiˇc J., Manning C.D., Pyysalo S., Schuster S., Tyers F., Zeman D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of the 12th Language Resources and Evaluation Conference. 2020. Marseille, France. European Language Resources Association. P. 4034–4043.

Pollard C., Sag I.A. Head-Driven Phrase Structure Grammar. University of Chicago Press, 1994.

Saba, W.S. Stochastic LLMs do not understand language: towards symbolic, explainable and ontologically based LLMs. In International Conference on Conceptual Modeling. Cham: Springer Nature Switzerland, 2023. P. 3–19.

Starace G., Papakostas K., Choenni R., Panagiotopoulos A., Rosati M., Leidinger A., Shutova E. Probing LLMs for Joint Encoding of Linguistic Categories. arXiv preprint arXiv:2310.18696, 2023.