Unveiling the Syntax Within: Interpreting Grammar Embeddings in Meta's LLaMA Models

Pratim Chowdhary[*]
Department of Computer Science
Dartmouth College
cpratim.25@dartmouth.edu

Peter Chin
Department of Engineering
Thayer School of Engineering
pc@dartmouth.edu

Deepernab Chakrabarty
Department of Computer Science
Dartmouth College
deepernab@dartmouth.edu

Abstract

This paper investigates the mechanisms by which large language models (LLMs) encode grammatical knowledge, focusing on Meta's LLaMA models. By leveraging embedding vectors, we classify grammatically correct sentences and analyze the activation patterns of attention heads to identify their roles in processing specific grammatical structures. Furthermore, we explore the effects of selectively removing these attention heads, shedding light on how grammar is embedded within the model's architecture. Our findings aim to enhance the understanding of LLMs' linguistic capabilities and their internal organization of syntactic knowledge.

1 Introduction

2 Related Work

1. **Deciphering Stereotypes in Pre-Trained Language Models**

    - This paper goes into details about how stereotypes are encoded in the embeddings of pre-trained language models and how specific attention heads are responsible for encoding them.

2.

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2024. Do not distribute.

3 Methodology

4 Experiments

5 Discussion

6 Conclusion

References

References follow the acknowledgments in the camera-ready paper. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

A Appendix / supplemental material

Optionally include supplemental material (complete proofs, additional experiments and plots) in appendix.