Unveiling the Syntax Within: Interpreting Grammar Embeddings in Small-Scale LLMs

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

We investigate how grammatical knowledge is embedded within small language models (1-8B parameters), an understudied yet practically important class of LLMs. Through systematic analysis of Meta's Llama, Gemma, and Qwen model families using the BLiMP benchmark of linguistic minimal pairs, we examine grammatical representations across different architectures and parameter scales. Our findings reveal that grammatical knowledge scales non-uniformly, with certain model families demonstrating specific grammatical strengths independent of size. We observe complex grammatical phenomena showing non-linear scaling patterns, while others plateau quickly. Notably, smaller models from certain architectures sometimes outperform larger models from different families on specific grammatical tasks, suggesting architectural inductive biases significantly influence grammatical knowledge acquisition. This work provides insights into grammar representation in contemporary LLMs with implications for model design and enhancement of grammatical competence in resource-constrained deployment scenarios.

1 Introduction

Large language models (LLMs) have transformed natural language processing, demonstrating unprecedented capabilities across a spectrum of linguistic tasks—from simple question answering to complex reasoning, creative writing, and code generation [Brown et al., 2020, Chowdhery et al., 2022, Touvron et al., 2023b, Jiang et al., 2023]. These models, trained through self-supervised learning on vast corpora of text data, have increasingly approached human-like performance in generating coherent, contextually appropriate, and grammatically well-formed text, despite having no explicit grammatical rules programmed into their architecture. This emergent grammatical competence—arising purely from statistical patterns in training data—represents a fascinating case study in how linguistic knowledge can be acquired implicitly through exposure rather than explicit instruction.

The field has witnessed exponential growth in model size, from early transformer models with hundreds of millions of parameters [Vaswani et al., 2017] to modern giants like GPT-4 [OpenAI, 2023] that likely contain trillions of parameters. While these massive models have captured headlines with their impressive capabilities, a parallel revolution has been unfolding in the development of

*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.

smaller, more efficient models in the 1-8 billion parameter range. Models like Llama 2 [Touvron et al., 2023b], Gemma 7B [Jiang et al., 2023], and Qwen [Bai et al., 2023] have demonstrated remarkable performance despite their relatively modest size, making them particularly valuable for practical applications where computational efficiency and deployment costs are significant concerns.

These smaller LLMs offer a compelling balance between capability and efficiency, enabling deployment on consumer hardware, edge devices, and resource-constrained environments. Their reduced inference costs make them attractive for commercial applications, while their smaller memory footprint allows for fine-tuning and adaptation with more modest computational resources. Understanding how these models encode and represent grammatical knowledge is therefore not merely an academic exercise but has significant practical implications for developing more capable, efficient, and linguistically robust language technologies.

2 Related Work and Background

2.1 Grammatical Evaluation of Language Models

Research on grammatical capabilities in neural language models has evolved from early work on LSTM-based models [Linzen et al., 2016] to transformer-based architectures [Goldberg, 2019, Devlin et al., 2019]. Evaluation frameworks have progressed from simple subject-verb agreement tests to comprehensive benchmarks like CoLA [Warstadt et al., 2019] and BLiMP [Warstadt et al., 2020]. The BLiMP benchmark, comprising 67 minimal pair tests across 12 linguistic phenomena, has been particularly valuable for assessing grammatical competence. Recent studies have extended evaluation to larger models, revealing that while performance scales with size, challenges remain with complex hierarchical structures and garden-path sentences [Thrush et al., 2022, Qian et al., 2022].

2.2 Scale and Grammatical Competence

The relationship between model scale and grammatical abilities follows complex patterns beyond the general power-law scaling observed in language modeling [Kaplan et al., 2020]. Research suggests that rare grammatical constructions require disproportionately more training data [Wei et al., 2021], and certain phenomena show nonlinear improvements at specific parameter thresholds [Zhang et al., 2023]. Studies on compositional generalization indicate that scaling alone may not capture human-like grammatical productivity without architectural innovations [Hu and Daumé III, 2020]. Beyond raw parameter count, specific architectural components significantly impact grammatical knowledge acquisition, with attention mechanisms crucial for long-distance dependencies and feed-forward networks encoding categorical information [Patel and Pavlick, 2022].

2.3 Probing Language Models for Linguistic Knowledge

Researchers have developed various probing techniques to understand how linguistic knowledge is represented within model parameters. Structural probes have revealed that models implicitly encode hierarchical syntactic structures [Hewitt and Manning, 2019], with syntactic information appearing in earlier layers and semantic information in later ones [Tenney et al., 2019]. Transformer-based models exhibit emergent symbolic manipulation capabilities resembling discrete linguistic rules [Manning et al., 2020], with individual neurons specializing in specific linguistic features [Geva et al., 2021]. Attention patterns often correspond to syntactic dependencies [Lazaridou et al., 2018], with different attention heads specializing in distinct linguistic phenomena [Clark et al., 2019].

2.4 Small LLMs and Comparative Studies

The proliferation of open-weight models like OPT [Zhang et al., 2022], Llama [Touvron et al., 2023a], Gemma [Jiang et al., 2023], and Qwen [Bai et al., 2023] has enabled more systematic analyses of how capabilities scale and how architectural choices affect performance. Studies show that smaller models with innovative architectures can outperform larger ones [Jiang et al., 2023], and data quality may be as important as scale for robust grammatical representations [Bai et al., 2023]. Comparative analyses across architectures remain limited but suggest significant variations in performance even at similar parameter scales [Talmor et al., 2020, Zhao et al., 2023]. Recent frameworks for quantifying evaluation uncertainties [Roberts et al., 2023] and structured evaluation approaches [Xia et al., 2023]

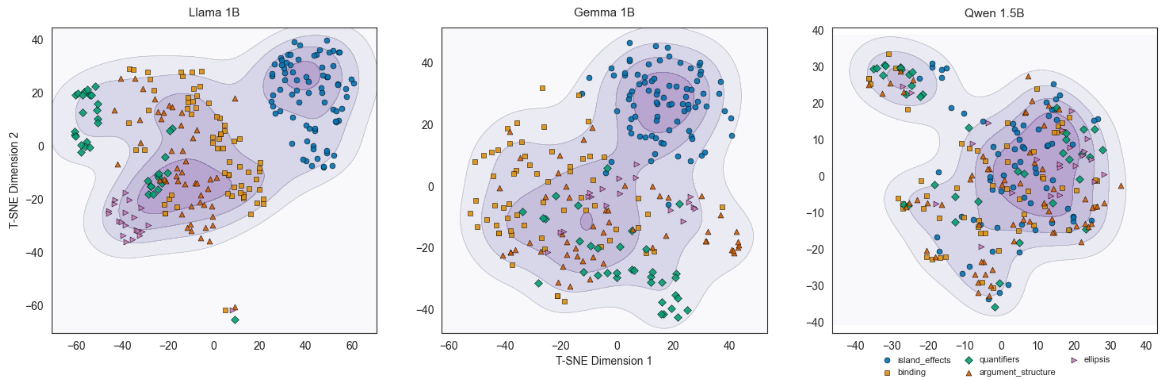

Figure 1: T-SNE visualizations of grammatical representations across three 1B-scale language models. Points represent ungrammatical sentences colored by linguistic category. Contours indicate density of representations, revealing how different architectures organize grammatical knowledge.

have revealed that architectural design choices impact specific grammatical phenomena differently, suggesting that model architectures encode grammatical knowledge in fundamentally different ways.

3 Methodology

3.1 Evaluating Grammatical Knowledge Across Linguistic Categories

Our evaluation framework divides grammatical phenomena into five key linguistic categories: island effects, quantifiers, ellipsis, argument structure, and binding constraints. This categorization allows us to assess whether grammatical knowledge is acquired uniformly or if certain phenomena are represented differently across model architectures and scales.

To visualize how these grammatical categories are embedded within model representations, we applied t-SNE dimensionality reduction to final-layer activations from ungrammatical sentences. Figure 1 shows these representations for three 1B-scale models. The visualizations reveal distinct organizational patterns, with varying degrees of cluster separation suggesting that models encode grammatical phenomena differently. Llama 1B forms more distinct category clusters, while Qwen 1.5B shows more integrated representations across categories.

3.2 Experimental Protocol for Grammatical Assessment

We evaluate grammatical knowledge in LLMs through a systematically designed protocol that captures both explicit and implicit grammatical understanding. Our methodology encompasses two complementary evaluation paradigms:

- **Binary Classification**: Models evaluate the grammaticality of individual sentences through explicit prompted judgment. This tests the model's ability to directly assess syntactic well-formedness.
- **Minimal Pair Discrimination**: Models select between grammatical and ungrammatical alternatives that differ minimally in structure. This approach aligns with psycholinguistic methodologies for investigating human grammatical intuitions.

The prompt templates for each evaluation paradigm are formalized as:

$$\mathcal{P}_{\text{binary}} = \text{“Is this sentence grammatically correct/incorrect?: } [S]\text{”} \quad (1)$$

$$\mathcal{P}_{\text{pair}} = \text{“Which sentence is grammatically better? (A) } [S_1] \text{ (B) } [S_2]\text{”} \quad (2)$$

Our analysis extends beyond surface-level responses to probe the internal representations formed during grammatical processing. For each sentence evaluation, we extract activation vectors from all

n transformer layers, enabling a fine-grained examination of how grammatical knowledge emerges and propagates through the network hierarchy:

$$\mathbf{A}_s = \{\mathbf{a}_{s,l_1}, \mathbf{a}_{s,l_2}, \ldots, \mathbf{a}_{s,l_n}\} \tag{3}$$

where $\mathbf{A}_s$ represents the complete set of layer-wise activations for sentence s. These activation patterns serve as a neural signature of the model's grammatical processing.

To quantify and compare grammatical sensitivity across different architectural layers, we compute the Euclidean distance between activation patterns elicited by grammatical (g) and ungrammatical (u) sentence pairs:

$$\text{GramDist}(l_i) = \frac{1}{|P|} \sum_{(g,u)\in P} \|\mathbf{a}_{g,l_i} - \mathbf{a}_{u,l_i}\|_2 \tag{4}$$

where P denotes the set of all grammatical/ungrammatical sentence pairs in our evaluation corpus. This metric captures the degree to which each layer's representations distinguish between well-formed and ill-formed syntactic structures.

Figure **??** illustrates the layer-wise grammatical sensitivity across different model families, revealing architectural variations in grammatical knowledge acquisition.

4 Experiments

5 Discussion

6 Conclusion

References

Jinze Bai, Shuai Lv, Sheng Peng, Yida Wang, Xingjian Zhang, Ziyue Yang, Beilei Yang, Haotian Gong, Zhiyu Fu, Kongming Liu, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

Yoav Goldberg. Assessing bert's syntactic abilities. In *arXiv preprint arXiv:1901.05287*, 2019.

John Hewitt and Christopher D Manning. 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)*, pages 4129–4138, 2019.

Weihs Hu and Hal Daumé III. Systematic evaluation of causal discovery in visual model based reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12578–12590, 2020.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Charles Blundell. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *arXiv preprint arXiv:1705.11192*, 2018.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Krishna Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. *arXiv preprint arXiv:2210.02539*, 2022.

Peng Qian, Tahsina Huang, Reza Firoozi, Zhiyuan Wang, Qiyang Zhou, Eric Wong, Kevin Chen, Shaopeng Pan, Zhou Yu, Yang Xiang, et al. Limitations of language models in arithmetic and symbolic reasoning. *arXiv preprint arXiv:2208.05051*, 2022.

Adam Roberts, Albert Webson, Colin Larson, Leo Gao, Niket Tandon, Kai-Wei Tai, Hyung Won Chung, Colin Raffel, and Gaurav Mishra. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*, 2023.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. olmpics-on what language model pre-training captures. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 743–758. MIT Press, 2020.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, 2019.

Tristan Thrush, Sanjay Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 625–641. MIT Press, 2019.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Jerry Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Frequency effects on syntactic rule learning in transformers. *arXiv preprint arXiv:2109.07020*, 2021.

Jiacheng Xia, Songbo Li, Haozhao Xu, Danny Chen, Yang Liu, Bill Cohen, and Leyang Zhang. Structured prompting: Scaling in-context learning to 1,000 examples. *arXiv preprint arXiv:2212.06713*, 2023.

Hugh Zhang, Amy Webb, Saujas Petryk, Yiheng Han, Jason Lei, and Chelsea Finn. Language modeling with reduced spurious correlations. *arXiv preprint arXiv:2306.01708*, 2023.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

A Appendix / supplemental material

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