Evaluating distillation methods for data-efficient syntax learning

Anonymous ACL submission

Abstract

Developing more data-efficient training approaches depends on a better understanding of inductive biases. In this work, we hypothesize that the structural information encoded in a transformer's attention matrices is key to acquiring syntax because attention captures relationships between words – a crucial part of syntax. Under this hypothesis, we would expect that inductive biases targeting attention should selectively improve data-efficiency on syntactic benchmarks. We use knowledge distillation (KD) as a methodological lens to test this hypothesis, comparing conventional KD through output logits against KD through attention matrices. Using GPT-2 as our teacher model, we train student models on datasets ranging from 10K to 5M sentences and evaluate them on both syntactic benchmarks and general language modeling tasks. Surprisingly, we find that while logit-based KD drastically improves data-efficiency across all metrics, attention-based KD offers minimal benefits even for syntactic tasks. This suggests that logits already effectively supervise syntactic information, challenging assumptions about how syntax is represented in transformers and informing more targeted approaches to data-efficient training.

1 Introduction

Modern language models successfully capture many aspects of human linguistic competence, from the fundamentals of grammar (Warstadt et al., 2020; Linzen and Baroni, 2021; Hu et al., 2024) to more sophisticated uses of world knowledge (Ivanova et al., 2024; Yamakoshi et al., 2023). However, they achieve these capabilities only after training on vastly more data than human children receive during language acquisition (Frank, 2023), motivating research into *inductive biases* (Warstadt et al., 2023) – predispositions that guide learning toward particular solutions with less data. These biases include architectural modifications (Sartran et al., 2022), curriculum learning strategies (Martinez et al., 2023), and specialized weight initialization techniques (Bencomo et al., 2025).

In this paper, we use knowledge distillation (KD) to study which aspects of a model's learned representations are most critical for scaffolding particular linguistic capabilities. We focus specifically on learning syntax – an ability long theorized to require strong (innate) biases (Chomsky, 1965; McCoy et al., 2020). Previous research has shown that syntactic information is encoded in the attention mechanism of transformer models (Clark et al., 2019), and that constraining these attention matrices can serve as an effective inductive bias for syntax (Nguyen et al., 2020; Qian et al., 2021; Yoshida and Oseki, 2022; Sartran et al., 2022). These studies raise an intriguing hypothesis: if attention matrices are the locus of syntactic knowledge, then distillation specifically targeting these representations ought to transfer syntactic abilities just as efficiently, or more efficiently, than conventional distillation through output logits.

To test this hypothesis, we performed a controlled experiment using a pretrained GPT-2 model (Radford et al., 2019) as the teacher, and trained student models of identical architecture on datasets ranging from 10K to 5M sentences. Our contributions are twofold. First, we demonstrate that conventional distillation through an additional supervision signal on logits can drastically reduce the amount of data required for learning syntax, reaching teacher-level performance with only 1M sentences of training data. Second, more surprisingly, we show that attention-based KD offers limited benefits for syntactic tasks despite prior evidence that these matrices encode crucial structural information. Our work illustrates how knowledge distillation can serve as a powerful analytical tool for understanding which aspects of a model's representations are effective for achieving data-efficiency with respect to specific linguistic capabilities.

2 Related Work

2.1 Knowledge distillation

Knowledge distillation (KD) consists of three main approaches (Gou et al., 2021): response-based KD, which aligns the output distributions of teacher and student models; feature-based KD, which matches internal representations to transfer detailed computational patterns; and relation-based KD, which preserves relational structures across multiple samples. In this work, we employ both response-based KD through logits and feature-based KD through attention to investigate their relative effectiveness for transferring syntactic knowledge.

While KD was initially developed for model compression, its applications have been expanded in several directions. For example, Furlanello et al. (2018) demonstrated that distilling knowledge to a student of identical architecture can actually improve performance. Others have used KD to facilitate transfer between architecturally different models (Kuncoro et al., 2019, 2020; Abnar et al., 2020), showing that inductive biases from specialized architectures can be distilled into more general ones. Finally, recent work has explored KD for data-efficient training, using ensembles of teacher models to improve student performance on limited data (Timiryasov and Tastet, 2023; Samuel, 2023; Yam and Paek, 2024). Our approach maintains architectural consistency between teacher and student, and uses a single pre-trained model as the teacher, in order to isolate the effects of different distillation mechanisms on syntactic competencies.

2.2 How transformers represent syntax

Understanding how transformers capture syntactic structure has been a central question in interpretability research. Numerous studies have identified attention matrices as repositories of syntactic information, with certain attention heads specializing in tracking specific syntactic relations (Clark et al., 2019; Vig and Belinkov, 2019; Htut et al., 2019) and incorporating explicit syntactic guidance into attention patterns can improve performance on syntactic tasks (Strubell et al., 2018; Sachan et al., 2021; Bugliarello and Okazaki, 2019; Wang et al., 2019b; Bai et al., 2021; Chen et al., 2024).

Recent work has also investigated the data requirements for acquiring syntactic knowledge, with some studies finding that pre-training on small, developmentally plausible corpora can lead to syntax acquisition with the right inductive biases (Warstadt et al., 2023; Huebner et al., 2021). However, the precise mechanisms through which transformers acquire syntactic knowledge, and the relative contributions of different elements of the architecture, remain open questions.

3 Approach

We ask whether distillation through attention provides a stronger inductive bias for syntax acquisition compared to conventional distillation through logits. To investigate this question, we conducted controlled experiments using the GPT-2 small architecture (Radford et al., 2019) for both the teacher and student models. The teacher model was a fully pre-trained GPT-2, while the student models were trained from scratch on different subsets of the BabyLM dataset (Warstadt et al., 2023), ranging from 10K to 5M sentences. By varying the dataset size, we assessed how different distillation methods affect data efficiency. All results reported are averages across three random seeds. Complete training details are provided in Appendix A.

3.1 Distillation via logits

We first established the baseline effectiveness of conventional KD through output distributions. Following Kim and Rush (2016), we implemented word-level KD where the student model learns to match the teacher's output probability distributions. Let $P_t(w|w_{<i})$ and $P_s(w|w_{<i})$ be the conditional probability of the word w at the i-th token calculated by the teacher and the student model respectively. The auxiliary loss for distillation $\mathcal{L}_{\text{logits}}$ for each sentence with length N was defined as

$$\mathcal{L}_{\text{logits}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{w\in V} P_t(w|w_{<i}) \log P_s(w|w_{<i}),$$

where V is the vocabulary. This formulation is equivalent to calculating the forward KL divergence between teacher and student distributions at each token position and taking the average. This auxiliary loss was then added to the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ with a coefficient α controlling the strength of distillation:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{logits}}.$$

Based on preliminary experiments testing different values of α, we found that $\alpha = 10$ led to optimal performance and fixed it at this value for all logit-based distillation experiments.

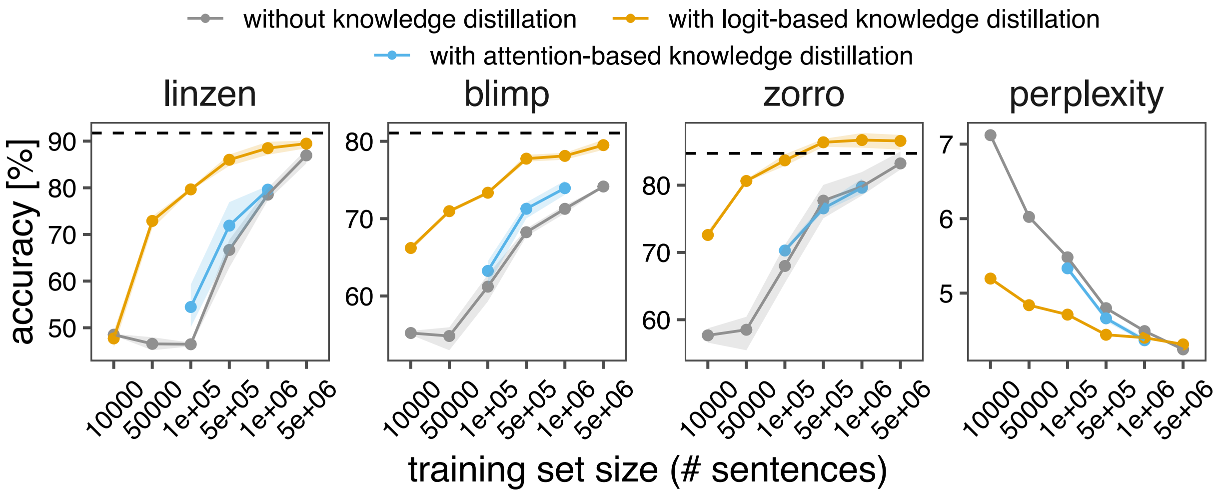

Figure 1: Performance of the students trained on datasets with different sizes. Linzen, BLiMP, and Zorro are targeted syntactic evaluations, while perplexity quantifies general language modeling performance. Ribbons show the bootstrapped 95% CI. Dashed lines indicate the performance of the teacher.

3.2 Distillation via attention

To test our hypothesis that attention matrices might provide a stronger inductive bias for syntax acquisition, we implemented feature-based KD targeting the attention mechanisms directly. We calculated the auxiliary loss $\mathcal{L}_{\text{attn}}$ as the mean squared error between the attention matrices of the teacher and the student. Let $A_t(l, h)$ and $A_s(l, h)$ be the attention matrices of the head h at layer l calculated by the teacher and the student model, respectively.

$$\mathcal{L}_{\text{attn}} = \frac{1}{L}\frac{1}{H}\sum_{l=1}^{L}\sum_{h=1}^{H}\text{MSE}(A_t(l,h) - A_s(l,h)),$$

where L and H are the number of layers and heads. As with logit-based distillation, this auxiliary loss was added to the cross-entropy loss with a coefficient α, which we set to 1 based on preliminary experiments.

3.3 Evaluation

To test our hypothesis about the relative effectiveness of different distillation approaches for syntax acquisition, we evaluated models on both syntactic benchmarks and a conventional language modeling metric. If attention matrices encode critical syntactic information not fully captured in output distributions, then attention-based distillation should show selective advantages on syntactic tasks, especially when training data is limited. For syntactic evaluation, we used three datasets based on minimal pairs:

- **Linzen** (Linzen et al., 2016; Gulordava et al., 2018) tests subject-verb agreement across various syntactic constructions.
- **BLiMP** (Warstadt et al., 2020) tests 67 distinct tasks across 12 syntactic phenomena.
- **Zorro** (Huebner et al., 2021): tests basic syntactic tasks that align with the developmental nature of our training data.

For each item in these benchmarks, we computed the log probability of both sentences and counted the model as correct if it assigns a higher probability to the grammatically acceptable variant. To ensure we capture overall language modeling capability (beyond syntax), we also measured perplexity on the BabyLM test split. This dual evaluation allows us to distinguish between general improvements in language modeling and selective enhancements in syntactic competence, helping to determine whether different distillation methods provide domain-specific inductive biases or general learning benefits.

4 Results

Before testing the effects of KD on syntactic performance, we first check to make sure that each KD approach achieves what it is intended to do. As shown in S1, this is indeed the case: logit-based KD enables the student model to have a much lower KL divergence from the teacher model, and attention-based KD enables the student model to have a much more similar attention pattern to the teacher model. Now that we have established that each KD method is effective for its training objective, we turn to our main question: how does each KD method affect the linguistic abilities of the student models?

4.1 Logit-based KD improves data efficiency

Figure 1 shows the performance of students trained with and without KD via logits across varying dataset sizes. KD resulted in substantial improvements on both syntactic benchmarks and perplexity. With just 1M sentences (approx. 10M tokens), the students approached the performance of the teacher that was trained on billions of tokens, demonstrating the remarkable data efficiency of KD.

The impact of logit-based KD was particularly pronounced with smaller datasets, where inductive biases are most crucial. For models trained on just 50K-100K sentences, KD provided a >20% boost in performance on the Linzen benchmark, elevating models from chance-level performance (50%). This indicates that KD can serve as a powerful inductive bias that enables syntax acquisition even with very limited data.

Interestingly, some students outperformed the teacher on the Zorro benchmark. This may reflect the domain alignment between the student's training data and the benchmark, which uses the vocabulary from the BabyLM dataset, whereas the teacher's training data was a more general Internet-based corpus. This result suggests that distillation can combine the teacher's knowledge and the domain-specific property of the student's training data.

4.2 Attention-based KD has a limited effect

Contrary to our hypothesis that attention matrices provide a stronger inductive bias for syntax acquisition, Figure 1 shows that attention-based KD offered limited benefits compared to logit-based KD, even though it leads to better alignment in attention S1. This pattern held consistently across all dataset sizes tested, suggesting that the syntactic information encoded in attention matrices may not provide substantial advantages beyond what is already captured in output distributions.

To determine whether attention-based KD selectively benefits particular aspects of syntax, we performed fine-grained evaluations across individual tasks and grammatical phenomena. Figure S2 breaks down performance by tasks, and Figure S3 by phenomena, in the BLiMP benchmark. Despite considerable variation in the teacher's performance across these tasks and phenomena, the relative performance pattern of different distillation approaches remained remarkably consistent. Similar patterns were observed for the Zorro benchmark (Figure S4).

5 Discussion

Our results reveal a striking contrast in the ability to improve data-efficiency among different KD methods. While KD via logits enabled student models to achieve teacher-level syntactic performance with just 1M sentences, KD via attention matrices – despite their capacity to encode syntactic structures – offered only marginal benefits.

One explanation is that logit-based KD indirectly aligns attention patterns, making explicit attention distillation redundant (Wu et al., 2024). A preliminary analysis supports this hypothesis: when both KD methods are combined, performance remains similar to logit-based KD alone (Figure S5), suggesting no unique contribution from attention-based KD. This indicates that output distributions may provide sufficient signal to scaffold data-efficient syntax learning, suggesting that syntax might be encoded redundantly throughout the network rather than being localized primarily in attention patterns.

One key advantage of KD is that it requires minimal assumptions about the specific form of inductive biases. In fact, our results demonstrate that strong syntactic performance can be achieved without relying on explicit grammatical rules. On the other hand, KD-based approaches present certain challenges. KD can be computationally intensive, requiring forward passes through the teacher model for the entire training dataset, and the inductive biases transferred via KD are less interpretable than those from explicit grammar-based approaches (Sartran et al., 2022).

Our findings highlight how feature-based KD can serve as a powerful analytical tool to investigate which features are most critical for specific capabilities. Effective distillation through a particular feature suggests that it contains information that works as an inductive bias for the target capability. Our results suggest that the information contained in attention matrices was not a strong enough inductive bias for syntax acquisition, but future work must systematically compare different feature-based KD methods to better understand how different linguistic competencies are encoded within transformer representations.

Limitations

Our evaluation focused specifically on syntactic benchmarks, motivated by previous work showing that attention matrices encode syntactic information and that syntactically-guided attention constraints serve as effective inductive biases. While this targeted approach allowed us to directly address questions about syntax acquisition, it limits the generalizability of our findings to other linguistic competencies. Different aspects of linguistic knowledge may be encoded preferentially in different components of transformer architectures, and distillation methods might show varying effectiveness across other linguistic domains, from semantics and pragmatics to discourse representation. Further work should systematically compare feature-based KD methods across a broader range of linguistic capabilities to develop a more complete understanding of knowledge representation in these models.

Future work should evaluate attention-based KD on a broader range of benchmarks spanning diverse capabilities, such as SuperGLUE (Wang et al., 2019a) for language understanding and EWOK (Ivanova et al., 2024) for world knowledge. A more comprehensive evaluation would allow researchers to determine whether the relative efficacy of different distillation methods varies across linguistic domains. It's possible that attention-based distillation might provide stronger benefits for capabilities other than syntax, such as long-range semantic dependencies or pragmatic reasoning.

Additionally, our experiments used a single pretrained model (GPT-2) as the teacher. Exploring different teacher architectures and scales would help determine the generalizability of our findings across different model families and capabilities. Finally, our exploration of feature-based distillation was limited to attention matrices; future work could investigate other internal representations such as hidden states, feed-forward network activations, or combinations of these features.

Ethics Statement

All datasets (BabyLM, Linzen, BLiMP, and Zorro) and the model (GPT-2) used in this paper were employed according to their intended usage. BabyLM consists of the following publicly available datasets (Warstadt et al., 2023):

- CHILDES[1] (MacWhinney, 2000)
- British National Corpus[2] (Consortium, 2007)
- Children's Book Test (Hill et al., 2016)
- Children's Stories Text Corpus (Bensaid et al., 2021)
- Project Gutenberg (Gerlach and Font-Clos, 2020)
- OpenSubtitles (Lison and Tiedemann, 2016)
- QED (Abdelali et al., 2014)
- Wikipedia
- Simple English Wikipedia
- Switchboard Corpus (Godfrey et al., 1992)

While we utilized knowledge distillation (KD) to distill the inductive biases required for data-efficient syntax learning, KD can also transfer the biases embedded in the teacher. When training student models using KD, we need to consider the biases of the teacher as well as those in the training dataset.

Acknowledgments

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Samira Abnar, Mostafa Dehghani, and Willem Zuidema. 2020. Transferring inductive biases through knowledge distillation. *arXiv preprint arXiv:2006.00555*.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, and 1 others. 2024. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, pages 929–947.

[1]CC BY-NC-SA 3.0 License
[2]BNC License

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-BERT: Improving pre-trained transformers with syntax trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020.

Gianluca Bencomo, Max Gupta, Ioana Marinescu, R Thomas McCoy, and Thomas L Griffiths. 2025. Teasing apart architecture and initial weights as sources of inductive bias in neural networks. *arXiv preprint arXiv:2502.20237*.

Eden Bensaid, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. 2021. Fairytailor: A multimodal generative framework for storytelling. *arXiv preprint arXiv:2108.04324*.

Emanuele Bugliarello and Naoaki Okazaki. 2019. Enhancing machine translation with dependency-aware self-attention. *arXiv preprint arXiv:1909.03149*.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L Leavitt, and Naomi Saphra. 2024. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *The Twelfth International Conference on Learning Representations*.

Noam Chomsky. 1965. Aspects of the theory of syntax.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. 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, Florence, Italy. Association for Computational Linguistics.

BNC Consortium. 2007. The british national corpus, xml edition.

Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016*.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in BERT track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121. Publisher: Proceedings of the National Academy of Sciences.

Philip A Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning*, pages 624–646.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, and 1 others. 2024. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Adhiguna Kuncoro, Chris Dyer, Laura Rimell, Stephen Clark, and Phil Blunsom. 2019. Scalable syntax-aware language models using knowledge distillation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3472–3484, Florence, Italy. Association for Computational Linguistics.

Adhiguna Kuncoro, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, and Phil Blunsom. 2020. Syntactic structure distillation pretraining for bidirectional encoders. *Transactions of the Association for Computational Linguistics*, 8:776–794.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

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

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB–curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127.

R Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Xuan-Phi Nguyen, Shafiq Joty, Steven Hoi, and Richard Socher. 2020. Tree-structured attention with hierarchical accumulation. In *International Conference on Learning Representations*.

Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernandez Astudillo. 2021. Structural guidance for transformer language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3735–3745, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Devendra Sachan, Yuhao Zhang, Peng Qi, and William L Hamilton. 2021. Do syntax trees help pre-trained transformers extract information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2647–2661.

David Samuel. 2023. Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 221–237, Singapore. Association for Computational Linguistics.

Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. 2022. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew Mccallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby Llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 279–289, Singapore. Association for Computational Linguistics.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019b. Tree Transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and 1 others. 2023. Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Cindy Wu, Ekdeep Singh Lubana, Bruno Kacper Mlodozeniec, Robert Kirk, and David Krueger. 2024. What mechanisms does knowledge distillation distill? In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pages 60–75. PMLR.

Hong Meng Yam and Nathan Paek. 2024. Teaching tiny minds: Exploring methods to enhance knowledge distillation for small language models. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 302–307.

Takateru Yamakoshi, James L McClelland, Adele E Goldberg, and Robert D Hawkins. 2023. Causal interventions expose implicit situation models for commonsense language understanding. *Findings of the Association for Computational Linguistics: ACL 2023*.

Ryo Yoshida and Yohei Oseki. 2022. Composition, attention, or both? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5822–5834, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Training details

Table S1 shows hyperparameters used in our experiments. The BabyLM preprocessing pipeline[3] was used to clean the dataset. Since the dataset has one sentence per line, we used the number of sentences as the measure of dataset size rather than the number of words or tokens. All train runs had the same number of training steps (156,250 steps) except for those for the largest dataset size (5,000,000 sentences). We used a linear warm-up for 1% of the total number of training steps.

We used Hugging Face `transformers` (version 4.45.2; Apache License 2.0) (Wolf et al., 2020) and PyTorch (version 2.4.1; BSD-style license [4]) (Ansel et al., 2024) to train and evaluate models. Experiments took approximately 750 GPU hours with NVIDIA RTX A6000 GPUs.

`n_layers`	12
`n_heads`	12
`hidden_size`	768
`intermediate_size`	3072
`max # tokens`	128
`batch size`	32
`learning rate`	0.0002

Table S1: Hyperparameters

[3] https://github.com/babylm/babylm_data_preprocessing

[4] https://github.com/pytorch/pytorch/blob/main/LICENSE

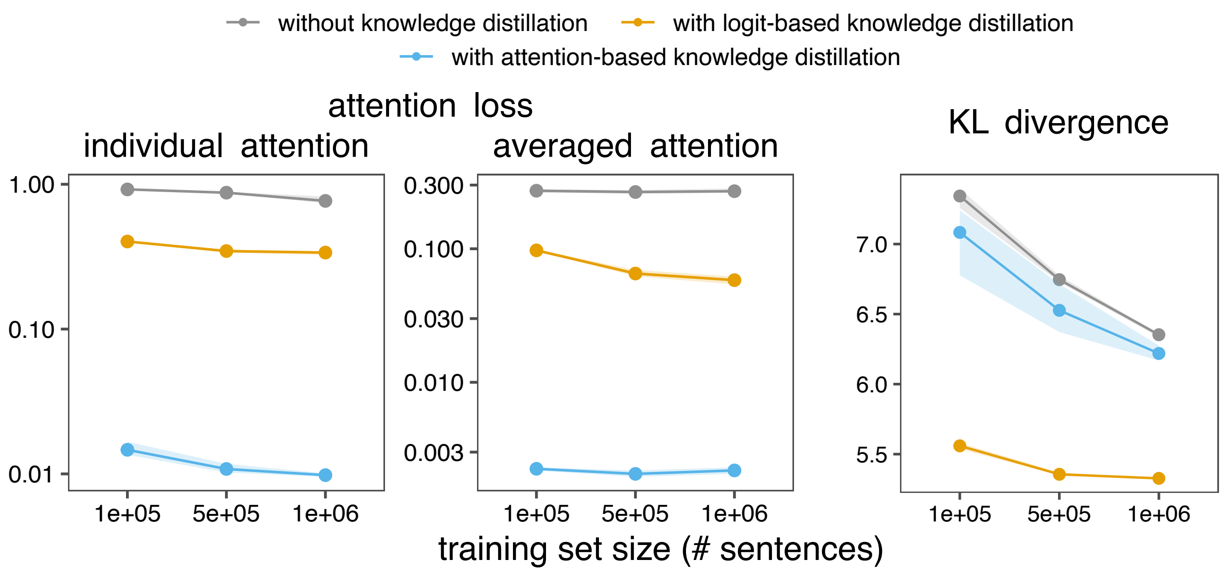

Figure S1: Auxiliary losses evaluated on the BLiMP dataset. We randomly selected 3 items from each task (3*67=201 in total). Unlike attention-based knowledge distillation, logit-based knowledge distillation does not align the internal computations, which leaves the possibility that similar attention patterns are implemented in both the teacher and the student by different attention heads. To account for this, we calculated the loss using the attention matrices averaged across layers and heads (middle), in addition to the loss used in training (left) as described in 3.2. Y-axis of the left two panels are on the log scale.

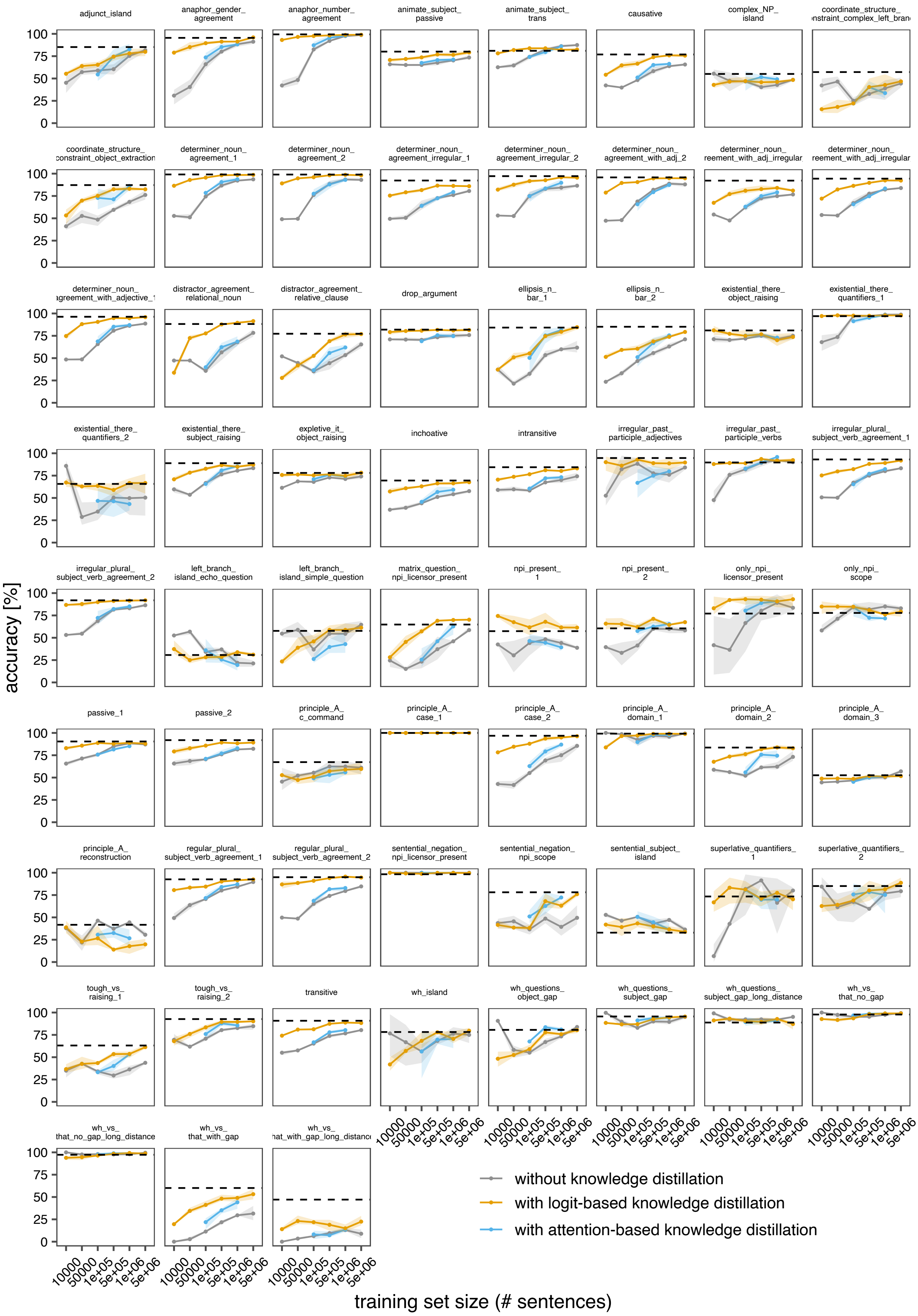

Figure S2: Performance on BLiMP split into tasks. Ribbons show the bootstrapped 95% CI.

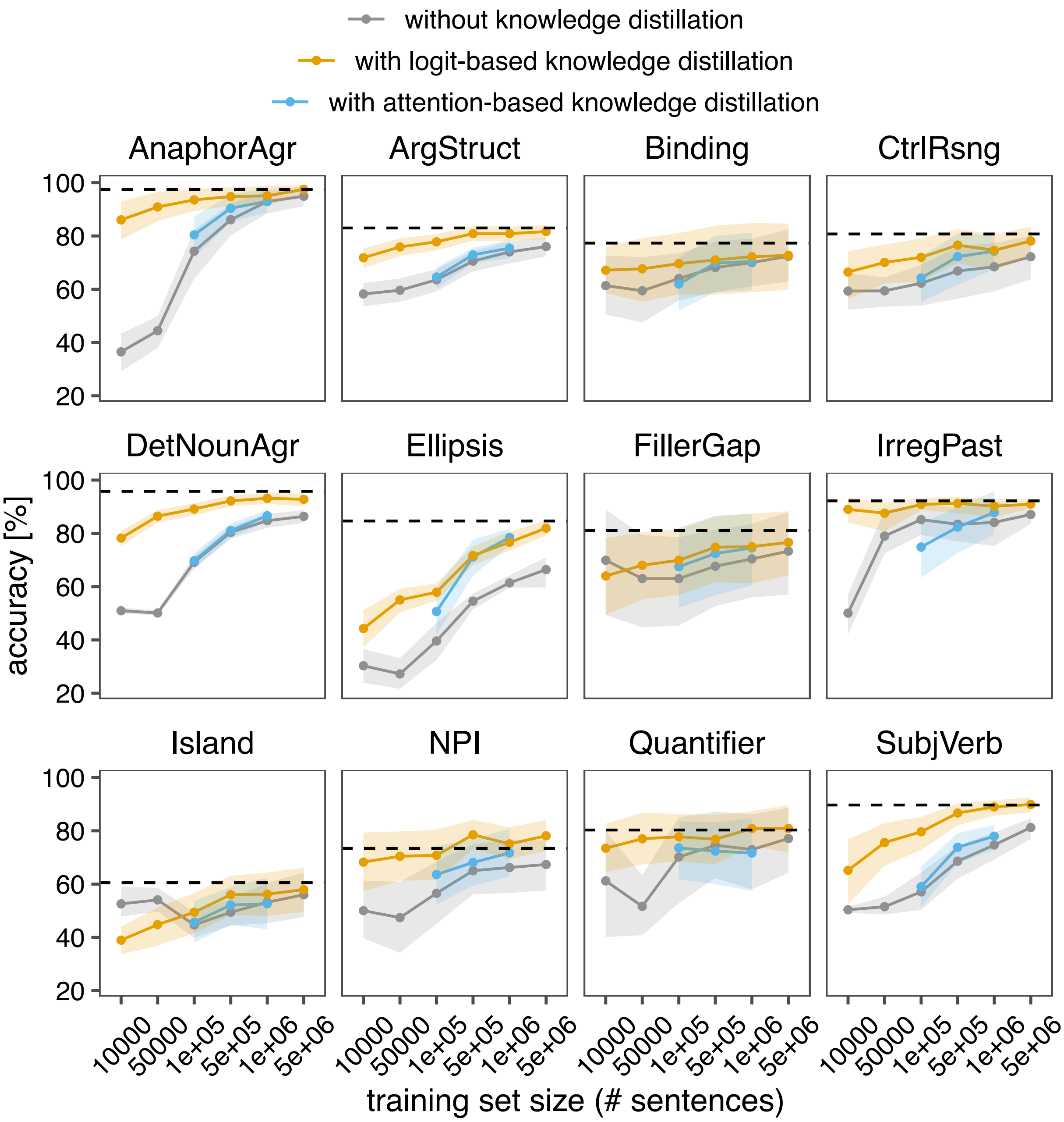

Figure S3: Performance on BLiMP split into phenomena. Ribbons show the bootstrapped 95% CI.

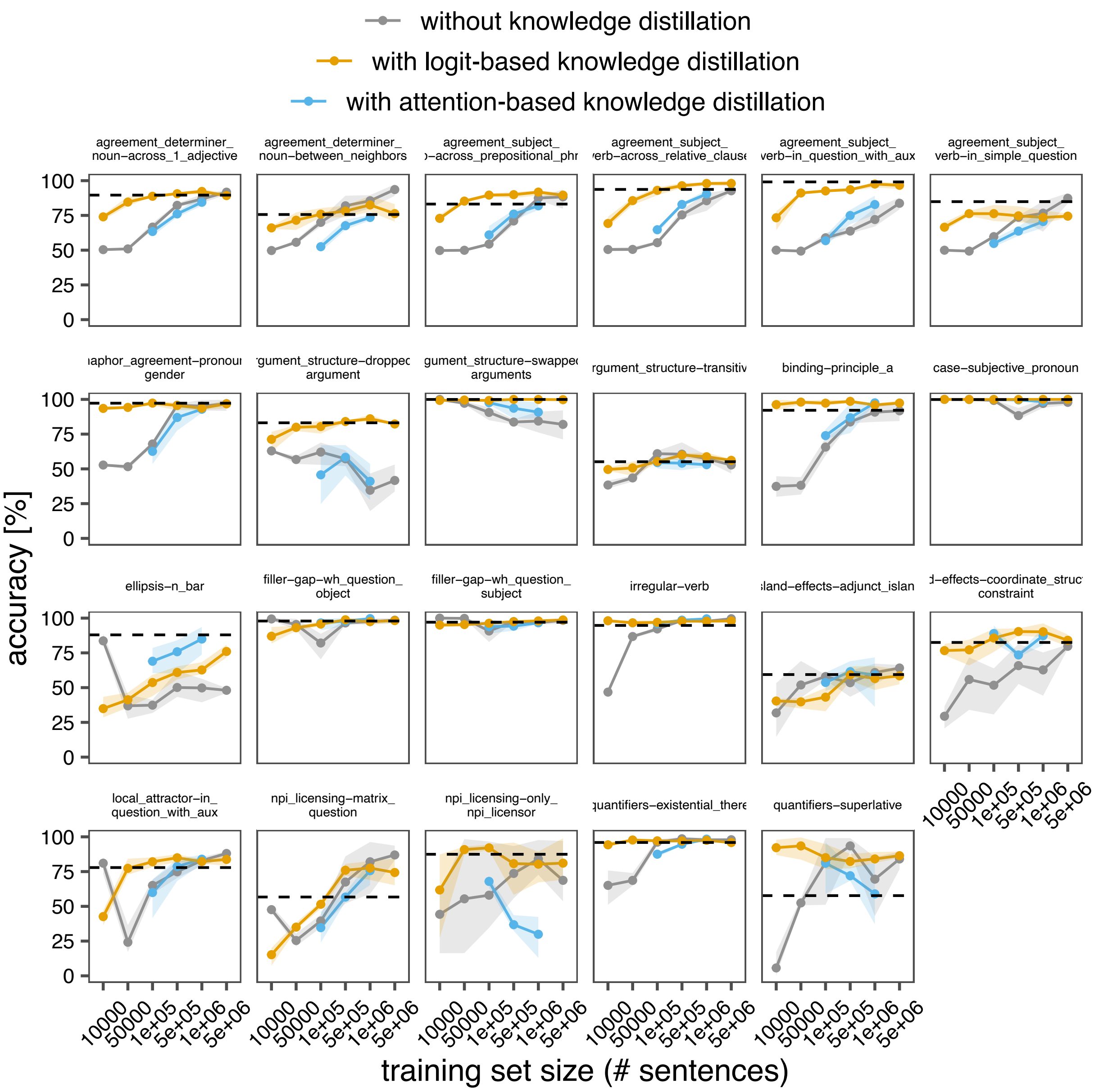

Figure S4: Performance on Zorro split into tasks. Ribbons show the bootstrapped 95% CI.

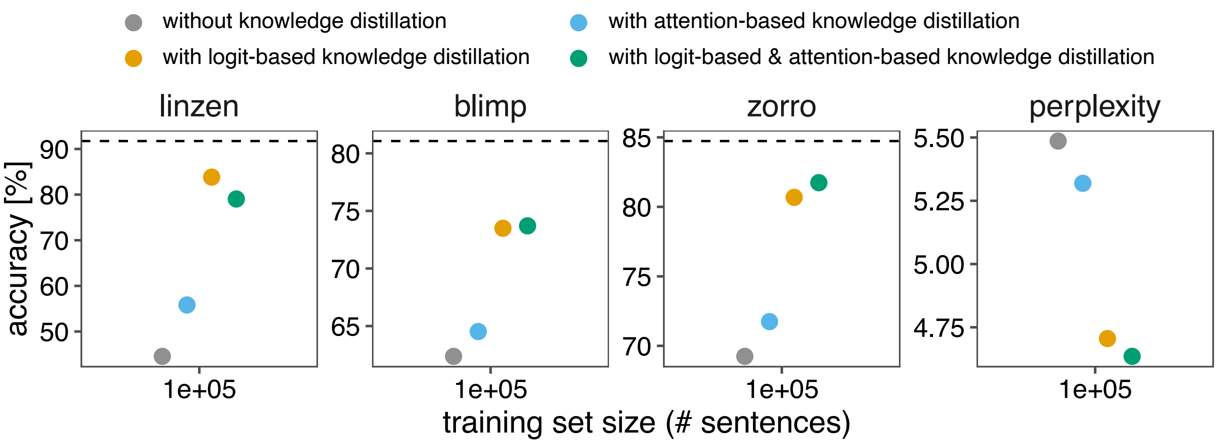

Figure S5: Preliminary analysis showing little unique effects of KD through attention matrices.