# Very-Long-Distance Dependency Capturing Evaluation via Language Modeling Based on Gender Consistency

Hongfei Xu[1], Zhuofei Liang[1], Josef van Genabith[2], Deyi Xiong[3], Hongying Zan[1], Qiuhui Liu[4], and Tengxun Zhang[1,5](✉)

[1] Zhengzhou University, Zhengzhou 450001, Henan, China
iehyzan@zzu.edu.cn, ztx313@foxmail.com

[2] German Research Center for Artificial Intelligence and Saarland Informatics Campus, 66123 Saarbrücken, Germany
Josef.van_Genabith@dfki.de

[3] College of Intelligence and Computing, Tianjin University, Tianjin 300072, China
dyxiong@tju.edu.cn

[4] China Mobile Online Services, Zhengzhou 450001, Henan, China

[5] Henan Branch Agricultural Bank of China, Zhengzhou 450001, Henan, China

**Abstract.** Capturing Long-Distance Dependencies (LDDs) is crucial for NLP applications. However, the longest relation evaluated in early studies is only around 50 words, which is not sufficient to evaluate a model's ability in capturing Very-Long-Distance (VLD) dependencies. Recent work on capturing LDDs either is affected by the instruction-following ability of language models or requires training on synthetic tasks unrelated to natural languages. In this paper, we present an approach to automatically constructing LDD test instances (as opposed to training examples) for any distance by mentioning an antecedent with singular number and a specific grammatical gender at the start of the first sentence, building the first sentence of arbitrary length by sampling plural nouns, and asking the pre-trained language model to predict a singular pronoun with the correct gender at the start of the next sentence. We evaluate the performance of LLMs and neural language models with different settings.

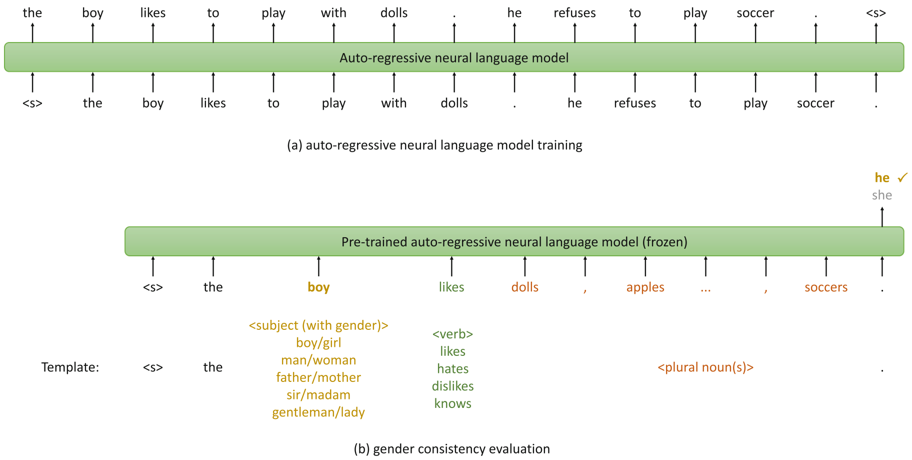
**Keywords:** Long-distance dependency · Language model · Grammatical gender consistency

## 1 Introduction

Capturing Long-Distance Dependencies (LDDs) is crucial for the good performance of NLP applications [1–9] based on Large Language Models (LLMs) [10–20].

[21] builds the contrastive *Lingeval97* test set for the subject-verb agreement task by swapping the grammatical number of a verb to introduce an agreement

X.-L. Mao et al. (Eds.): NLPCC 2025, LNAI 16105, pp. 27–38, 2026.
https://doi.org/10.1007/978-981-95-3352-7_3


the boy likes to play with dolls . he refuses to play soccer . <s>
Auto-regressive neural language model
<s> the boy likes to play with dolls . he refuses to play soccer .
he ✓
she
Pre-trained auto-regressive neural language model (frozen)
<s> the boy likes dolls , apples ... , soccers .
Template:
<s>
the
<subject (with gender)>
boy/girl
man/woman
father/mother
sir/madam
gentleman/lady
<verb>
likes
hates
dislikes
knows
<plural noun(s)>
.

(a) auto-regressive neural language model training

(b) gender consistency evaluation

**Fig. 1.** Auto-regressive neural language model training and gender consistency evaluation.

error. But quite a large proportion of distances in *Lingeval97* are within 15 tokens, and cannot support the Very-Long-Distance (VLD) dependency evaluation (e.g., more than 100 tokens). Recently, [22] evaluate the performance of question answering, summarization, and code completion on long documents. [23] ask LLMs to retrieve random facts inserted into a long document. [24] train the model to generate the value tokens of corresponding key tokens in long key-value pair input sequences. These evaluations either depend on the instruction-following ability of LLMs, which can significantly impact performance (as studied in Sect. 3.1) [22, 23], or require to train models on task-specific synthetic datasets which do not have any relation to natural languages [24]. They also lack a focus on the real distances between dependencies despite offering long inputs.

We present a template-based method to build large-scale sentences and test (**not train**) LDDs based on grammatical gender consistency in language modeling. **Our evaluation does not rely on instruction following, but tests on grammatical sentences and focuses on the dependency distance.**

Our main contributions are as follows:

- We present an automatic LDD evaluation method based on **purely language modeling** by evaluating the grammatical gender agreement between singular words marked for grammatical gender, controlling VLD length by adding plural nouns after the verb.
- We evaluate the performance of LLMs, test the effects of various settings and provide valuable empirical results and references for the design of neural language models.

## 2 VLD Evaluation via Grammatical Gender Consistency

In natural languages, the grammatical gender of the pronoun in the following sentences should be consistent with the co-referring subject in a previous sentence, as shown in Fig. 1 (a). We build test sentences where we control the grammatical gender of singular subjects (e.g., “the boy”/“the girl”, etc.) and singular co-referring pronouns (“he” or “she”), and control the dependency length using only plural intervening nouns, as shown in Fig. 1(b). The language model shall assign a higher probability to a co-referring pronoun of the same gender as the subject than to a pronoun with another grammatical gender when decoding the first token (i.e., the singular pronoun) of the next sentence. **As none of plural nouns (by design) can function as antecedent to s/he, the prediction is expected to be consistent with the grammatical gender of the subject.**

We measure the VLD distance by the number of intervening plural nouns, and build a balanced test set across the two classes. The subject set $S$ contains 10 singular words with clear grammatical gender ({“boy”, “girl”, “man”, “woman”, “father”, “mother”, “sir”, “madam”, “gentleman”, “lady”}), the verb set $V$ has 4 verbs ({“likes”, “hates”, “dislikes”, “knows”}), while the plural noun set $N$ has 6204 plural nouns (for animals (noun.animal), artifacts (noun.artifact), food (noun.food) and plants (noun.plant) extracted from Wordnet). **Our choice of intervening plural nouns between singular antecedent (reference) and singular anaphor is to make sure that s/he is an unambiguous binary choice and to be able to vary dependency length via intervening plural nouns, where none of the plural nouns (by design) can function as antecedent to s/he.** Human filtering of plural nouns can in principle make the generated test set more reasonable, but here we want to avoid such human efforts to make the test set construction method fully automatic and scalable.

For each distance $d$, we automatically build 5k test instances, and this already leads to 1.28M test instances in total for 256 distances, and we think this is sufficient to provide a reliable evaluation result. But at the same time it is also easy to produce more test instances and for longer distances with the method.

**All test instances are grammatically correct. We compared average per token loss between the Lingeval97 (4.2) and our synthetic dataset (4.5) using a Transformer LM. This shows that it is reasonable to use the synthetic test set for the evaluation at varying distances.**

## 3 Evaluation

In addition to Large Language Models (LLMs), we also trained Transformer and RNN models to examine the effects of various settings on VLD.

As training on long sequences is likely to be crucial for the model’s ability in capturing VLD dependency, we trained autoregressive neural language models on the document-split version of the English News Crawl dataset from WMT [25]. The datasets were lower cased, and tokenized into subwords by the BART

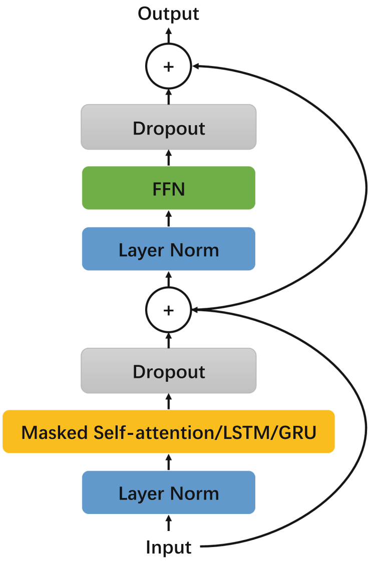


**Fig. 2.** The computation graph of the model layer.

tokenizer [26] using Byte-Pair Encoding models [27,28]. We concatenated consecutive sentences in documents into one sequence until adding the next sentence increases the sequence length to more than $j$ tokens, where $j$ was defaulted to 768.

Following the settings of [29], we used a batch size of 25k tokens. All the base, deep and big models were trained for 100k steps unless otherwise stated. The number of warm-up steps was set to 8k.

As for the model architectures, we employed the pre-norm computation order of the Transformer, which computes the layer normalization before the multi-head attention/position-wise feed-forward layer, and applies dropout and residual connection in the end, and used the absolute positional encoding unless indicated otherwise. For RNN models, we replaced the self-attention sub-layer in the Transformer by the LSTM/GRU sub-layers instead of purely stacking LSTM/GRU layers to minimize differences between the models. The computation graph of the model layer is shown in Fig. 2. We used the base setting by default, and the original absolute positional encoding for the Transformer.

In Figs. 3 and 5, 6 and 7, the x-axis and y-axis are the number of intervening plural nouns and the VLD accuracy respectively.

### 3.1 Performance of LLMs

For LLMs, we evaluated the performance of LLaMa 3.1–8B [30], Qwen 2.5–7B [31], GLM 4 9B 0414 [32] and Mamba - 2.8B [33] on the VLD test set. We evaluated the model in both language modeling (LM) and instruction following

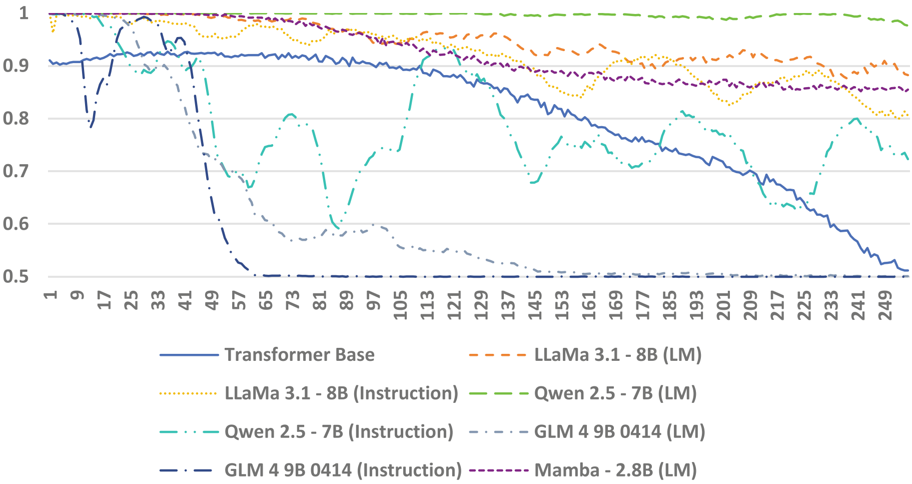


**Fig. 3.** Results of LLMs.

(Instruction) for LLaMa, Qwen and GLM. We only reported the LM evaluation performance of Mamba as its instruction performance is significantly poorer. For language modeling, we regard the model as a language model, feed the test input sequence into the model, and get the prediction probabilities of "he" and "she" at the last position. For instruction following, we ask the model to continue the input sentence starting with either "he" or "she", providing 4 examples as demonstrations.

Results in Fig. 3 show that **evaluation with instruction following leads to generally worse performances than with language modeling for both LLaMa and Qwen on the same testset with increasing distances. Qwen outperforms LLaMa in the LM evaluation but LLaMa performs better than Qwen in the Instruction evaluation, showing that instruction following can have a large impact on the long-distance evaluation.**

### 3.2 Self-attention vs. RNNs

We compare the performance of self-attentional Transformer, LSTM and GRU. When training on the dataset with a maximum sequence length of 768 tokens, we found that the LSTM model has difficulty in convergence even with residual connections. We conjecture the potential reason of the convergence issue is that LSTM may have problems when propagating gradients through the very long sequences.

We verified the convergence issue by training the LSTM model on the training data with a maximum sequence length of 256 (instead of 768) tokens and found that the model (LSTM_256) can converge. The per-token training loss averaged over $\sim$ 56M tokens reported during training is shown in Fig. 4.

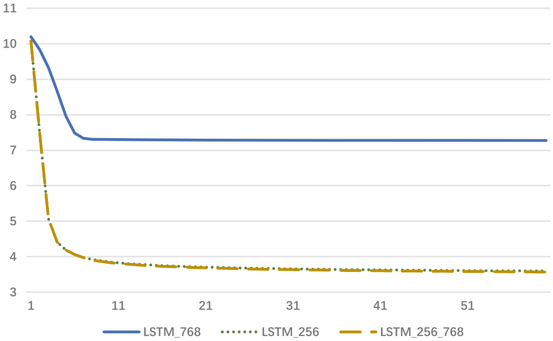
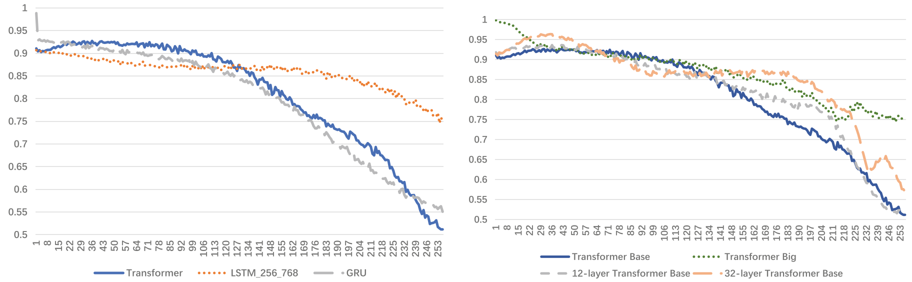


**Fig. 4.** Per-token training loss of LSTM models averaged over ∼56M tokens.



(a) Results of Transformer and RNNs. (b) Results of wide and deep Transformers.

**Fig. 5.** Results of architecture differences and wide/deep Transformers.

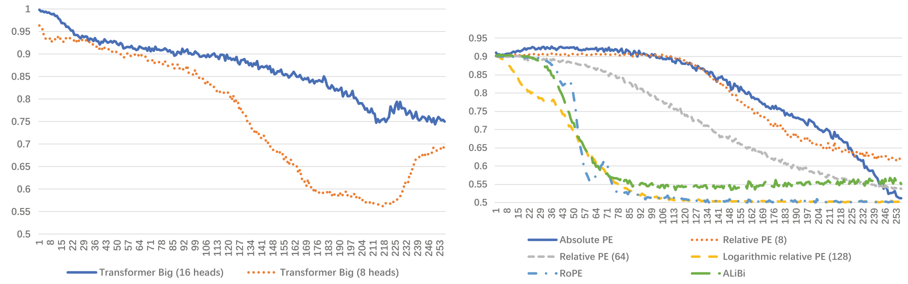
Figure 4 shows that the per-token training loss of LSTM_768 saturates at a high level while LSTM_256 achieves a much lower and reasonable training loss, demonstrating the convergence issue of LSTMs on very long sequences.

We employ a 2-stage training method for LSTM. The first stage trains the LSTM with a maximum sequence length of 256 tokens until the averaged per-token training loss is less than 4 (∼10k training steps), and the second stage continues the training on sequences with a maximum length of 768 for the remaining steps. GRU does not suffer from the convergence issue like LSTM when training on long sequences (of at most 768 tokens).

Results in Fig. 5a show that **LSTM unexpectedly outperforms Transformer for VLD** ($d > 128$), suggesting that recurrent architectures may still have the upper hand in VLD resolution worth further exploration.

### 3.3 Increasing Depth vs. Width

We test the effects of increasing model depth and width by comparing the Transformer Big (increasing the embedding dimension to 1024) with 12-layer and 32-layer deep Transformers (increasing the depth) trained for 100k steps. The



(a) Results of 8/16-head Transformer Big models.

(b) Results of positional encoding methods.

**Fig. 6.** Results of attention head numbers and position encoding methods.

12-layer Transformer has the same number of attention heads as the 6-layer Transformer Big, and the 32-layer model has a comparable amount of parameters as the Transformer Big.

Results in Fig. 5b show that **increasing the width is more effective in improving the performance on very long distances** ($d > 221$) **compared to increasing the depth**.

### 3.4 Effects of Attention Head Numbers

We verify the effects of the number of attention heads by comparing the 8-head Transformer Big model with the standard Transformer Big with 16 heads in each multi-head attention layer.

Results in Fig. 6a show that the 16-head Transformer Big model consistently outperforms the 8-head Transformer Big model in all distances, and it is important for wide models to have sufficient number of attention heads.

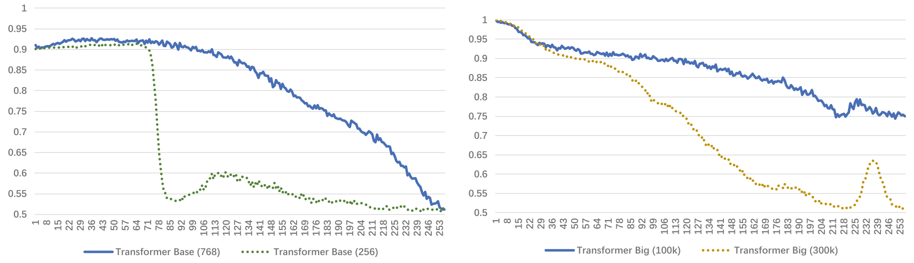
### 3.5 Absolute Positional Encoding vs. Relative Positional Encoding

We test the effects of absolute position encoding [29], and relative position encoding methods, including: [34] (Relative PE), [35] (Logarithmic relative PE), [36] (ALiBi) and [37] (RoPE). We explore different windows sizes (in parentheses) for Relative PE.

Results in Fig. 6b show that: 1) **absolute position encoding performs best for most distances**, and 2) Relative PE (8) performs better than absolute position encoding for very long distances ($d > 225$).

### 3.6 Effects of Training Sequence Lengths

To test the effects of training sequence lengths, we trained the Transformer Base models on sequences that contain at most 768 (Transformer Base (768)) and 256 (Transformer Base (256)) tokens respectively.



(a) Results of training sequence lengths. (b) Results of different training step numbers.

**Fig. 7.** Results of training sequence lengths and steps.

Results in Fig. 7a show that the Transformer Base (256) model consistently underperforms Transformer Base (768), and its performance drops sharply after a certain distance. This further proves that it is important to train on long sequences for good VLD performance.

### 3.7 Effects of Training Step Numbers

To test the effects of training step numbers on VLD dependency capturing ability, we trained the Transformer Big models for 100k (Transformer Big (100k)) and 300k (Transformer Big (300k)) steps respectively.

Results in Fig. 7b show that **Transformer Big** (100k) **outperforms Transformer Big** (300k) **for almost all distances**, especially with increasing distances, suggesting that **longer training may hinder VLD performance**.

## 4 Related Work

[38] test whether the verb form is consistent with subject number (singular or plural). [21] build the contrastive *Lingeval97* test set for the subject-verb agreement task by swapping the grammatical number of a verb. [39] present DistilLingEval based on machine-generated references. [40,41] explore automatic discourse phenomena tagging methods for context-aware machine translation. These studies cannot meet the requirements to assess the model's ability in capturing Very-Long-Distance (VLD) dependencies. Recently, [22] evaluate the performance of question answering, summarization, and code completion on long documents. [23] ask LLMs to retrieve random facts inserted into a long document. [24] ask the model to generate the values of corresponding keys in the long key-value pair input sequences. They either are affected by the instruction following ability of LLMs which can have a huge impact [22,23], or require to train models on synthetic tasks which have no relation to natural languages [24]. They also lack a focus on the real distances between dependencies despite offering long inputs.

[42] evaluate the performance of the Transformer, LSTM and CNN. [43] study the effects of phrase-level modeling, and [44] test Average Attention Network [45], Addition-subtraction Twin-gated Recurrent network [46], and MHPLSTM. But all these tests on the *Lingeval97* test set are not for LDDs, and lack an investigation on the impacts of different settings.

## 5 Conclusion

To mitigate the effects of varying instruction-following abilities of LLMs on Very-Long-Distance (VLD) dependency evaluation and synthetic evaluation unrelated to natural languages, we present a template-based approach to automatically constructing large testsets for arbitrary distances based on grammatical gender consistency.

We evaluate the VLD performance of LLMs and popular neural language models with different settings, and find that: 1) instruction following ability has a huge impact on the VLD evaluation, 2) 2-stage training can address the convergence issue of LSTM on very long sequences and lead to better performance than Transformer on very-long distances, 3) increasing the width improves VLD performance while increasing depth hampers performance, and 4) longer training tends to hamper the VLD performance.

**Acknowledgments.** We thank anonymous reviewers for their insightful comments. Hongfei Xu and Zhuofei Liang acknowledge the support of National Natural Science Foundation of China (Grant No. 62306284), China Postdoctoral Science Foundation (Grant No. 2023M743189), and the Natural Science Foundation of Henan Province (Grant No. 232300421386). Josef van Genabith is supported by BMBF (German Federal Ministry of Education and Research) via the project SocialWear (Grant No. 01IW20002). Deyi Xiong is supported by the National Key Research and Development Program of China (Grant No. 2023YFE0116400). Hongying Zan is supported by the Key Program of Natural Science Foundation of China (Grant No. U23A20316).

## References

1. Ouyang, L., et al.: Training language models to follow instructions with human feedback (2022)
2. Zhou, J., Ke, P., Qiu, X., Huang, M., Zhang, J.: Chatgpt: potential, prospects, and limitations. In: Frontiers of Information Technology & Electronic Engineering, pp. 1–6 (2023)
3. Katz, D.M., Bommarito, M.J., Gao, S., Arredondo, P.: Gpt-4 passes the bar exam. Available at SSRN 4389233 (2023)
4. Heck, M., et al.: ChatGPT for zero-shot dialogue state tracking: a solution or an opportunity? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 936–950. Association for Computational Linguistics, Toronto, Canada, July 2023. https://doi.org/10.18653/v1/2023.acl-short.81,

5. Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., Hershcovich, D.: Assessing cross-cultural alignment between ChatGPT and human societies: an empirical study. In: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pp. 53–67. Association for Computational Linguistics, Dubrovnik, Croatia, May 2023, https://aclanthology.org/2023.c3nlp-1.7
6. Lund, B.D., Wang, T.: Chatting about chatgpt: how may ai and gpt impact academia and libraries? Library Hi Tech News **40**(3), 26–29 (2023)
7. Lee, P., Bubeck, S., Petro, J.: Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. N. Engl. J. Med. **388**(13), 1233–1239 (2023)
8. Lecler, A., Duron, L., Soyer, P.: Revolutionizing radiology with gpt-based models: current applications, future possibilities and limitations of chatgpt. Diagn. Interv. Imaging **104**(6), 269–274 (2023)
9. Lin, J.C., Younessi, D.N., Kurapati, S.S., Tang, O.Y., Scott, I.U.: Comparison of gpt-3.5, gpt-4, and human user performance on a practice ophthalmology written examination. In: Eye, pp. 1–2 (2023)
10. Brown, T., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
11. Xue, L., et al.: mt5: a massively multilingual pre-trained text-to-text transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 483–498 (2021)
12. Zhang, Z., et al.: Cpm-2: large-scale cost-effective pre-trained language models. AI Open **2**, 216–224 (2021)
13. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
14. Wang, X., et al.: Self-consistency improves chain of thought reasoning in language models. In: The Eleventh International Conference on Learning Representations (2022)
15. Sanh, V., et al.: Multitask prompted training enables zero-shot task generalization. In: ICLR 2022-Tenth International Conference on Learning Representations (2022)
16. Wang, T., et al.: What language model architecture and pretraining objective works best for zero-shot generalization? In: International Conference on Machine Learning, pp. 22964–22984. PMLR (2022)
17. Xu, F.F., Alon, U., Neubig, G., Hellendoorn, V.J.: A systematic evaluation of large language models of code. In: Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming, pp. 1–10 (2022)
18. Thoppilan, R., et al.: Lamda: language models for dialog applications. arXiv preprint arXiv:2201.08239 (2022)
19. Chowdhery, A., et al.: Palm: scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022)
20. Touvron, H., et al.: Llama: open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
21. Sennrich, R.: How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 376–382. ACL, Valencia, Spain, April 2017, https://aclanthology.org/E17-2060

22. Bai, Y., et al.: LongBench: a bilingual, multitask benchmark for long context understanding. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3119–3137. ACL, Bangkok, Thailand, August 2024, https://aclanthology.org/2024.acl-long.172
23. Li, M., Zhang, S., Liu, Y., Chen, K.: Needlebench: can llms do retrieval and reasoning in 1 million context window? (2024), https://arxiv.org/abs/2407.11963
24. Arora, S., et al.: Zoology: measuring and improving recall in efficient language models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=LY3ukUANko
25. Kocmi, T., et al.: Findings of the 2022 conference on machine translation (WMT22). In: Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 1–45. ACL, Abu Dhabi, United Arab Emirates (Hybrid), December 2022, https://aclanthology.org/2022.wmt-1.1
26. Lewis, M., et al.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880. ACL, July 2020, https://doi.org/10.18653/v1/2020.acl-main.703
27. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1715–1725. ACL, Berlin, Germany, August 2016, https://doi.org/10.18653/v1/P16-1162
28. Kudo, T.: Subword regularization: Improving neural network translation models with multiple subword candidates. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 66–75. ACL, Melbourne, Australia, July 2018, https://doi.org/10.18653/v1/P18-1007
29. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 5998–6008. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
30. AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
31. Yang, Q.A., et al.: Qwen2.5 technical report. ArXiv **abs/2412.15115** (2024), https://arxiv.org/abs/2412.15115
32. GLM, T., et al.: Chatglm: a family of large language models from glm-130b to glm-4 all tools (2024)
33. Gu, A., Dao, T.: Mamba: linear-time sequence modeling with selective state spaces. ArXiv **abs/2312.00752** (2023), https://arxiv.org/abs/2312.00752
34. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 464–468. ACL, New Orleans, Louisiana, June 2018, https://doi.org/10.18653/v1/N18-2074
35. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020), http://jmlr.org/papers/v21/20-074.html
36. Press, O., Smith, N., Lewis, M.: Train short, test long: attention with linear biases enables input length extrapolation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=R8sQPpGCv0
37. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: Roformer: enhanced transformer with rotary position embedding (2022)

38. Linzen, T., Dupoux, E., Goldberg, Y.: Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Trans. Assoc. Comput. Linguist. **4**, 521–535 (2016). https://doi.org/10.1162/tacl_a_00115
39. Vamvas, J., Sennrich, R.: On the limits of minimal pairs in contrastive evaluation. In: Bastings, J., et al. (eds.) Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pp. 58–68. ACL, Punta Cana, Dominican Republic, November 2021, https://doi.org/10.18653/v1/2021.blackboxnlp-1.5
40. Wicks, R., Post, M.: Identifying context-dependent translations for evaluation set production. In: Koehn, P., Haddow, B., Kocmi, T., Monz, C. (eds.) Proceedings of the Eighth Conference on Machine Translation, pp. 452–467. ACL, Singapore, December 2023, https://doi.org/10.18653/v1/2023.wmt-1.42
41. Fernandes, P., Yin, K., Liu, E., Martins, A., Neubig, G.: When does translation require context? a data-driven, multilingual exploration. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 606–626. ACL, Toronto, Canada, July 2023, https://doi.org/10.18653/v1/2023.acl-long.36
42. Tang, G., Müller, M., Rios, A., Sennrich, R.: Why self-attention? a targeted evaluation of neural machine translation architectures. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4263–4272. ACL, Brussels, Belgium, Oct-Nov 2018, https://doi.org/10.18653/v1/D18-1458
43. Xu, H., van Genabith, J., Xiong, D., Liu, Q., Zhang, J.: Learning source phrase representations for neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 386–396. ACL, July 2020, https://doi.org/10.18653/v1/2020.acl-main.37
44. Xu, H., Liu, Q., van Genabith, J., Xiong, D., Zhang, M.: Multi-head highly parallelized LSTM decoder for neural machine translation. 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), pp. 273–282. ACL, August 2021, https://doi.org/10.18653/v1/2021.acl-long.23
45. Zhang, B., Xiong, D., Su, J.: Accelerating neural transformer via an average attention network. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1789–1798. ACL, Melbourne, Australia, July 2018, https://doi.org/10.18653/v1/P18-1166
46. Zhang, B., Xiong, D., Su, J., Lin, Q., Zhang, H.: Simplifying neural machine translation with addition-subtraction twin-gated recurrent networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4273–4283. ACL, Brussels, Belgium, Oct-Nov 2018, https://doi.org/10.18653/v1/D18-1459