Natural Context Drift Undermines the Natural Language Understanding of Large Language Models

Yulong Wu[1], **Viktor Schlegel**[1,2] **and Riza Batista-Navarro**[1]
[1] Department of Computer Science, University of Manchester, United Kingdom
[2] Imperial College London, Imperial Global Singapore
{yulong.wu, riza.batista}@manchester.ac.uk
v.schlegel@imperial.ac.uk

Abstract

How does the natural evolution of context paragraphs affect Question Answering (QA) in generative Large Language Models (LLMs)? To address this, we propose a framework for curating naturally evolved, human-edited variants of reading passages from contemporary QA benchmarks and for analysing LLM performance across a range of semantic similarity scores, which quantify how closely each variant aligns with Wikipedia content on the same article topic that the LLM saw during pretraining. Using this framework, we evaluate 6 QA datasets and 8 LLMs with publicly available training data. Our experiments reveal that LLM performance declines as reading passages naturally diverge from the versions encountered during pretraining–even when the question and all necessary information remains present at inference time. For instance, average accuracy on BoolQ drops by over 30% from the highest to lowest similarity bins. This finding suggests that natural text evolution may pose a significant challenge to the language understanding capabilities of fully open-source LLMs.

1 Introduction

Large Language Models (LLMs), pre-trained on massive web-scale corpora, have been shown to be highly effective at Question Answering (QA) over text passages (OpenAI et al., 2024; DeepSeek-AI et al., 2025b,a; Yang et al., 2025; OLMo et al., 2025), a task that has long served as a testbed for evaluating natural language understanding (NLU) (Chen, 2018). While such progress has fostered the perception that LLMs possess strong NLU capabilities already, much of this success may stem from shortcut exploitation rather than *genuine* language understanding (Wu et al., 2023; Levy et al., 2023; Bhuiya et al., 2024), as revealed by robustness probing–a commonly used evaluation methodology that goes beyond surface-level performance to assess authentic comprehension.

Existing QA robustness evaluation paradigms typically operate statically: a perturbation function is applied to the original benchmark test set, and model performance is then measured on the resulting challenge set (Wang et al., 2022). Differentiating from previous work, this paper offers a novel dynamic perspective on understanding the limitations of generative LLMs by asking: what happens when reading paragraphs continue to evolve and diverge from their appearance during pretraining? Such scenarios are prevalent in real-world applications, where textual data instances can evolve over time due to ongoing human edits, content updates, or contextual shifts (e.g., Wikipedia articles (Yang et al., 2017)), thus requiring LLMs to demonstrate genuine language understanding. To the best of our knowledge, however, there has been no systematic investigation of this phenomenon in QA.

To address this gap, we propose a framework to analyse how LLM performance changes as the reading paragraph semantically diverges from the content of its source in the model's training data. Among various instances of evolving text corpora, we focus on Wikipedia, as it serves as a primary source of reading passages in many widely used QA benchmark datasets (Rajpurkar et al., 2016, 2018; Clark et al., 2019; Wang, 2022; Ho et al., 2023), is commonly included in LLMs pretraining data (Soldaini et al., 2024; Zhao et al., 2025), and, most importantly, provides clearly documented revision histories that capture natural text evolution (Yang et al., 2017). This allows us to curate human-edited variants of passages that reflect natural changes over time. Our approach adopts a gradual perspective by computing continuous semantic similarity scores at the paragraph level and correlating them with LLMs QA accuracy.

Within the developed framework, we empirically evaluate six QA datasets and eight LLMs with fully open-source training corpora. Our study finds that, across models with different training corpora and

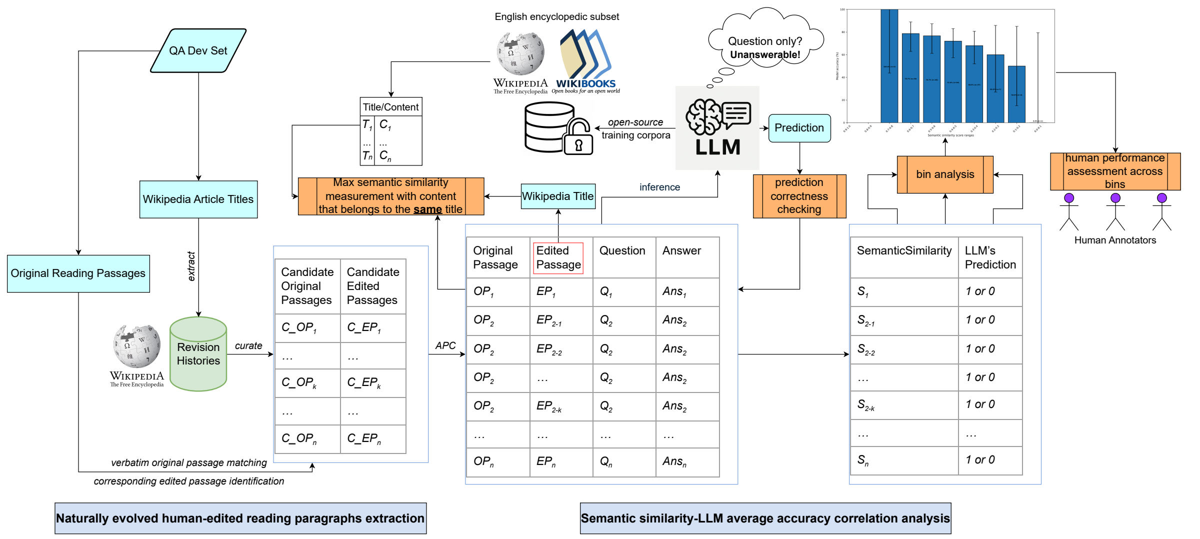


Figure 1: An overview of the analysis framework. Module *Naturally evolved human-edited reading paragraphs extraction* is adapted from (Wu et al., 2025) with minor modifications. APC: Answers Preserving Checking.

architectural configurations, as context paragraphs naturally evolve and become semantically distant from the Wikipedia content sharing the same article title seen during pretraining, the QA performance of LLMs generally deteriorates. In contrast, human annotators are less affected by such semantic drift and maintain relatively stable accuracy regardless of passage similarity, suggesting that the observed performance drop is specific to LLMs and not due to deficiencies in the edited passages themselves.

2 Methodology

In our framework (Figure 1), we extract revision histories of paragraphs from QA benchmarks, order them by semantic similarity to the version that appears in an LLM's training corpus, and correlate the LLMs' answer accuracy on those passages to the similarity thus obtained. The framework consists of two modules, described in detail below.

Naturally evolved human-edited reading paragraphs extraction. To obtain edited versions of original reading paragraphs from contemporary QA benchmark datasets that genuinely reflect real-world scenarios, we adopt the natural perturbation pipeline proposed by Wu et al. (2025), with two slight modifications: 1) we remove the constraint of retaining only candidate passage pairs where both paragraphs exceed 500 characters, allowing broader dataset applicability and preservation of diverse editing patterns; and 2) for the matched original passages with multiple occurrences, we retain all edited versions for each (see passage OP_2 in Figure 1 as an example) to support subsequent correlation analysis. Appendix A provides details on Answers Preserving Checking and data statistics.

Semantic similarity-LLM average accuracy correlation analysis. For each naturally evolved, human-edited reading paragraph and its associated question, we generate predictions using an LLM and label them as 1 (correct) or 0 (incorrect), based on a selected evaluation metric. We also collect predictions using the question alone to test whether the LLM already possesses parametric knowledge of the answer. Instances in which the LLM answers correctly without access to the passage are excluded, as they may call into question the paragraph's contribution to the answer (Glockner et al., 2025) [1]. Meanwhile, we extract English Wikipedia content from the LLM's training corpora that shares the same article title as the edited passage and compute the semantic similarity between them. The maximum similarity score is used as a proxy for how closely the passage resembles the training data. We then group the similarity scores into ten bins, compute the average LLM accuracy within each bin, and plot accuracy trends from highest to lowest similarity. Uncertainty for the result in each bin is estimated using the Wilson score interval with

[1]Appendix B shows that, within each generated dataset in Module *Naturally evolved human-edited reading paragraphs extraction*, the percentage of instances in which OLMo LLMs succeed on context-free QA, and are thus filtered out. The only exception is the BOOLQ dataset, where such instances are not filtered due to its straightforward yes/no answer format, which results in a significant proportion of questions being correctly answered or guessed by LLMs.

95% confidence ($z = 1.96$) (Wilson, 1927). Finally, to validate the observed trend, we assess human performance across the same bins.

3 Experiments Setup

Broadly, we address the following question: *How well do LLMs perform on QA as the reading paragraphs naturally evolve from the versions present in their training corpus?* To this end, we select QA benchmarks that feature context paragraphs from Wikipedia, whose edit histories allow us to trace naturally evolved versions of each paragraph, and evaluate LLMs with an open-source training corpus, as detailed below.

Datasets: We use the development set of six English QA datasets spanning extractive, yes/no, abstractive, cause-effect reasoning and multi-hop reasoning challenge: SQUAD 1.1 (Rajpurkar et al., 2016), SQUAD 2.0 (Rajpurkar et al., 2018), ADVERSARIALQA - D(ROBERTA) (Bartolo et al., 2020), BOOLQ (Clark et al., 2019), WIKIWHY (Ho et al., 2023) (version 1.2) and HOTPOTQA (Yang et al., 2018) (in the "distractor" setting). For HOTPOTQA, which includes paragraphs from multiple Wikipedia articles within a single context, we retain only the edits applied to the gold passages, i.e., those containing supporting facts that determine the answer, and disregard other distractors.

LLMs: We evaluate eight transparent instruction-tuned LLMs across six model families, all of which include Wikipedia in their publicly available training data: **OLMo** (OLMo-7B-0724-Instruct-hf) (Groeneveld et al., 2024), **OLMo 2** (OLMo-2-1124-7B-Instruct and OLMo-2-1124-13B-Instruct) (OLMo et al., 2025), **OLMoE** (OLMoE-1B-7B-0125-Instruct) (Muennighoff et al., 2025), LLM360's **AmberChat** (Liu et al., 2024b), **TinyLlama** (TinyLlama-1.1B-Chat-v1.0) (Zhang et al., 2024), Databricks' **Dolly** (dolly-v2-7b and dolly-v2-12b) (Conover et al., 2023). To isolate the potential impact of instruction tuning, we also conduct experiments on the base (non-aligned) versions of the same models. All LLMs are tested in a zero-shot setting; the inference prompts are provided in Appendix C. Model experimentation is carried out using HuggingFace's *Transformers* library (Wolf et al., 2020), the vLLM (Kwon et al., 2023), and two 80GB NVIDIA A100 GPUs. LLM prediction correctness is determined using Inclusion Match (IM), which considers a prediction correct if it includes any ground truth answer (Levy et al., 2023; Bhuiya et al., 2024). For WIKIWHY, due to its free-form format, correctness is based on semantic similarity (using the all-MiniLM-L6-v2 model) between the LLM's response and the ground truth answer; predictions scoring above 0.6 are considered correct.

Semantic Similarity Measure: We measure semantic-level textual similarity between the edited Wikipedia paragraph and the versions found in the Wikipedia subset of each LLM's training corpus–sourced from DOLMA V1.7[2] (Soldaini et al., 2024) for OLMo, REDPAJAMA V1[3] (Computer, 2023) for AmberChat, SLIMPAJAMA[4] (Soboleva et al., 2023) for TinyLlama and PILE[5] (Gao et al., 2020; Biderman et al., 2022) for Dolly, using a Sentence Transformers model all-MiniLM-L6-v2 (Reimers and Gurevych, 2019). In addition, we employ three alternative embedding models for the similarity measure to enhance the generalisability of our study: sentence-t5-base (Ni et al., 2022), all-mpnet-base-v2 and bge-small-en-v1.5 (Xiao et al., 2024). In measuring semantic similarity for HOTPOTQA, we consider only the paragraphs in the edited context whose corresponding Wikipedia article titles have matching content in the Wikipedia subset of an LLM's training corpus.

4 Results and Discussion

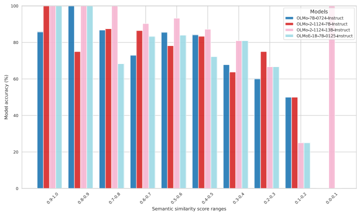
As a general trend, the QA performance of instruction-finetuned* OLMo *LLMs deteriorates as the reading paragraph semantically deviates from the training corpus. This is clearly visualized in Figure 2, where, across the evaluated benchmark datasets, model average accuracy generally declines as the reading paragraphs evolve and exhibit lower semantic similarity to the versions found in the Wikipedia subset of the training corpus. For example, on BOOLQ, the accuracy of OLMo-2-1124-7B-Instruct declines sharply from 71.1% in the highest similarity bin 0.9-1.0 to 12.5% in the lowest 0.1-0.2. Comparable drops are also observed, for instance, for OLMo-7B-0724-Instruct on SQUAD 2.0 and

[2] https://huggingface.co/datasets/allenai/dolma

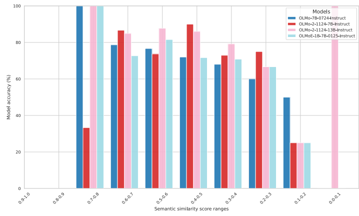
[3] https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T

[4] https://huggingface.co/datasets/cerebras/SlimPajama-627B

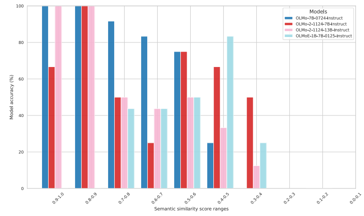
[5] As PILE is no longer officially hosted or distributed, we use a publicly accessible replication of its original Wikipedia component: https://github.com/noanabeshima/wikipedia-downloader.



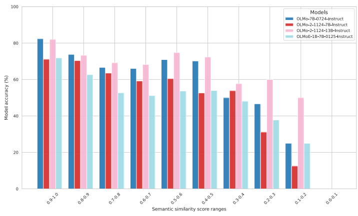
(a) SQuAD 1.1



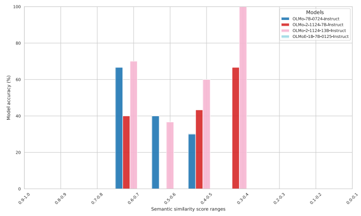
(b) SQuAD 2.0



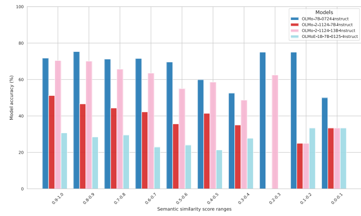
(c) D(RoBERTa)



(d) BoolQ

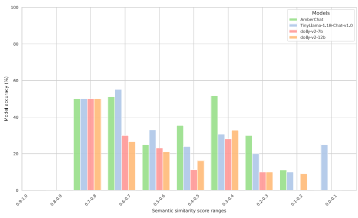


(e) WikiWhy

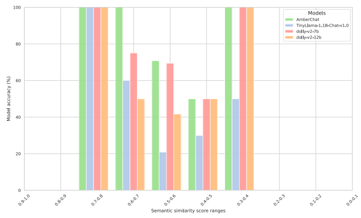


(f) HotpotQA

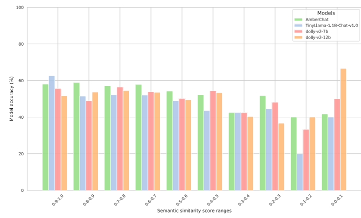
Figure 2: Average accuracy of the instruction-finetuned `OLMo` LLMs across ten semantic similarity bins.



(a) SQuAD 2.0



(b) WikiWhy



(c) HotpotQA

Figure 3: Average accuracy of other instruction-finetuned LLMs across ten semantic similarity bins.

WikiWhy, and for `OLMo-2-1124-13B-Instruct` on D(RoBERTa) and HotpotQA, illustrating the broader impact of natural context drift in diverse QA challenges. To further substantiate the observed trend, we perform a slope analysis as shown in Appendix D Figure 5, where linear regressions are fitted to each model's accuracy trajectory across semantic similarity bins. Aggregated across all tasks and models, the mean slope is 65.78 ± 41.55, with a mean Pearson correlation of 0.684 ± 0.291, highlighting a consistent and statistically grounded relationship between semantic divergence and performance degradation. Further, this downward trend is not exclusive to the `OLMo` family. As shown in Figure 3, ***a comparable decline in average performance across decreasing similarity ranges is likewise observed in other instruction-finetuned LLMs, including*** **`AmberChat`**, **`TinyLlama-1.1B-Chat-v1.0`** ***and*** **`dolly-v2-7b/12b`**, ***despite differences in model size, architecture, training corpora and procedure.*** To further investigate the generalisability of our findings, we conduct an ablation study using alternative embedding models for semantic similarity measurement (Figure 6 in Appendix E), evaluating the base pre-trained versions of the instruction-finetuned LLMs (Figures 7 and 8 in Appendix F) and prompting LLMs to perform chain-of-thought (CoT) reasoning (Wei et al., 2022) when generating responses (Figures 9 and 10 in Appendix G). ***Overall, the observed effects of context drift persist under these alternative settings, supporting the generalisability of our conclusions.***

We find from Figures 2, 3 and 5 that ***the impact of natural context drift is more pronounced in benchmarks that emphasize surface-form alignment, whereas tasks requiring deeper reasoning exhibit greater resilience.*** The downward trend is less pronounced for datasets requiring advanced reasoning, such as WikiWhy and HotpotQA. This may be because natural text evolution introduces diverse linguistic cues and contextual variations that activate broader reasoning mechanisms in LLMs, partially offsetting the impact of semantic drift. In contrast, surface-level QA tasks such as SQuAD rely more heavily on lexical or structural

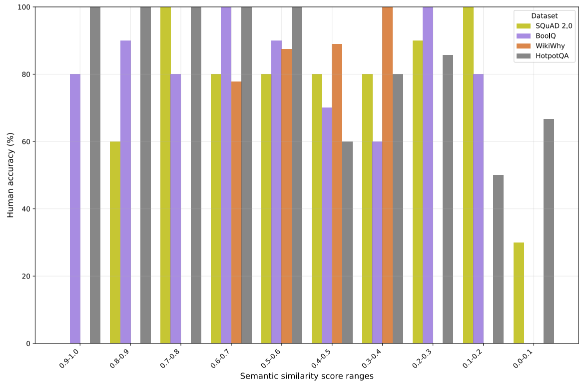


Figure 4: Accuracy of human annotators on QA tasks across semantic similarity bins.

cues (Schlegel et al., 2020; Wu et al., 2021) and therefore appear more sensitive to the text evolution. Dataset-level statistics support this contrast: BoolQ and SQuAD 2.0 exhibit quite steep average slopes (67.23 and 83.50, respectively) and strong Pearson correlations between average accuracy and semantic similarity (e.g., BoolQ's average correlation = 0.867 ± 0.064). Meanwhile, HotpotQA demonstrates a much shallower 23.94 average slope and lower correlation, indicating greater robustness to textual edits.

Unlike LLMs, human performance in reading comprehension is not influenced by deviations in measured semantic similarity. To determine whether the observed decline in LLMs' accuracy is due to reduced semantic similarity or the possibility that the edited reading paragraphs became degraded and unanswerable, we evaluate human performance across semantic similarity bins within the four investigated datasets. For each dataset, we randomly sample an equal number of edited QA instances from each bin, assign two annotators to label the answers, and involve a third annotator to resolve any disagreements. Details of the human annotation protocol are provided in Appendix H. As shown in Figure 4, across all QA benchmarks, human performance does not consistently decline with decreasing semantic similarity, supporting the conclusion that the degradation in LLM accuracy stems from semantic drift rather than a loss of question answerability.

There may be little concern regarding the impact of paragraph leakage from other data sources beyond Wikipedia. A natural question arises as to whether, given the vast scale of LLMs' training corpora, the edited versions of the reading paragraphs might also appear in sources beyond the Wikipedia subset we focus on, potentially affecting the findings. Therefore, using the infini-gram engine (Liu et al., 2024a), we aim to estimate as accurately as possible the percentage of edited reading paragraphs that appear verbatim in the complete training corpus of the evaluated LLMs across the six examined QA benchmarks, as shown in Table 1. Further details about the specific training corpora queried for each LLM are provided in Appendix I. As shown in Table 1, verbatim inclusion of the edited paragraphs in the LLMs' training corpora remains negligible across the board. With additional consideration that our analysis does not include all the edited reading paragraphs, we believe that the impact of these paragraphs appearing in other data sources may not be significant. Finally, we emphasise that our methodology focuses on measuring the semantic similarity between the edited reading paragraphs and the content from the same source, i.e., Wikipedia. Therefore, extending the analysis to other potential sources falls outside the scope of this paper, and we leave this for future investigation.

	SQuAD 1.1	SQuAD 2.0	D(RoBERTa)	BoolQ	WikiWhy	HotpotQA
OLMo-7B	1.59	3.94	1.33	4.30	1.52	-
OLMo-2-7B/13B	4.85	8.53	4.79	2.75	1.14	-
OLMoE-1B-7B	4.77	8.69	4.79	2.75	1.14	-
AmberChat&TinyLlama-1.1B	1.04	2.40	2.13	4.47	1.14	-
dolly-v2-7b/12b	0.18	1.07	-	0.40	0.76	-

Table 1: Percentage (%) of edited reading paragraphs that appear verbatim in the training corpora of the evaluated LLMs.

5 Conclusion

We introduce a novel methodology for examining how real-world natural context evolution affects the language understanding of LLMs when it deviates semantically from their training data. Leveraging Wikipedia revision histories, we curate naturally human-edited variants of benchmark reading passages, compute their semantic similarity to versions present in the models' training corpus, and correlate these similarity scores with model performance. Our empirical findings show that, while natural text evolution has little to no effect on human QA performance, recent fully open-source LLMs generally exhibit a consistent downward trend in accuracy as semantic similarity decreases. We hope this study contributes to the growing body of research focused on understanding and addressing the limitations of LLMs in real-world, evolving textual contexts.

Limitations

Our work has several limitations. First, we focus exclusively on the QA task; extending these findings to other downstream tasks remains an open avenue for future research. Second, we evaluate only transparent LLMs with fully open-access training corpora. While this constraint is necessary for our methodology, some of these models still lag behind state-of-the-art proprietary LLMs in performance. Applying our framework to such proprietary models and investigating the generalisability of our conclusions to them would be a valuable extension, although it would require access to their training data, which is currently unavailable. Finally, we note that the findings of this paper may be influenced by the accuracy of semantic similarity measurement, other implicit forms of benchmark leakage, and the evaluation metrics used for LLMs, all of which warrant further investigation in future work.

Ethical Considerations

All datasets, the extracted edited versions of the reading passages, and LLMs used in this work are publicly available, used consistently with their intended purpose and under the permitted license. A very small proportion of the human-edited paragraphs may contain offensive content, as they originate from reverted Wikipedia revisions that were intentionally introduced to damage articles. We retain these cases to enable a comprehensive study of the potential impact of natural text evolution on LLMs' QA performance.

Acknowledgements

This research is part of the IN-CYPHER programme and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. The authors thank the anonymous reviewers of the ACL Rolling Review (May 2025) for their insightful feedback and valuable suggestions. We are also grateful to the Computational Shared Facility at The University of Manchester for providing the resources that enabled our experiments. Yulong Wu is supported by the Kilburn Scholarship from the Department of Computer Science, The University of Manchester.

References

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Neeladri Bhuiya, Viktor Schlegel, and Stefan Winkler. 2024. Seemingly plausible distractors in multi-hop reasoning: Are large language models attentive readers? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2514–2528, Miami, Florida, USA. Association for Computational Linguistics.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the pile. *Preprint*, arXiv:2201.07311.

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Max Glockner, Xiang Jiang, Leonardo F. R. Ribeiro, Iryna Gurevych, and Markus Dreyer. 2025. Neoqa: Evidence-based question answering with generated news events. *Preprint*, arXiv:2505.05949.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. 2023. Wikiwhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Mosh Levy, Shauli Ravfogel, and Yoav Goldberg. 2023. Guiding LLM to fool itself: Automatically manipulating machine reading comprehension shortcut triggers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8495–8505, Singapore. Association for Computational Linguistics.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024a. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, and 8 others. 2024b. LLM360: Towards fully transparent open-source LLMs. In *First Conference on Language Modeling*.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, and 5 others. 2025. Olmoe: Open mixture-of-experts language models. *Preprint*, arXiv:2409.02060.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. `https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama`.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai

Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and improve robustness in NLP models: A survey. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4569–4586, Seattle, United States. Association for Computational Linguistics.

Zhen Wang. 2022. Modern question answering datasets and benchmarks: A survey. *Preprint*, arXiv:2206.15030.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2021. Is the understanding of explicit discourse relations required in machine reading comprehension? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3565–3579, Online. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2023. Are machine reading comprehension systems robust to context paraphrasing? In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 184–196, Nusa Dua, Bali. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2025. Pay attention to real world perturbations! natural robustness evaluation in machine reading comprehension. *Preprint*, arXiv:2502.16523.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in Wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010, Copenhagen, Denmark. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Answer Preservation Check and Data Statistics

For answer preservation checking, in the extractive QA setting (including the extractive question set for HotpotQA), we ensure that at least one (or all) of the ground truth answers can still be found in the edited passage. For BoolQ, we manually inspect the generated edited test set and remove instances where the edited passage contains fewer than 56 characters. We also check WikiWhy, but no filtering is applied. Table 2 presents the statistics of the extracted data for each QA dataset.

Dataset	Titles (ext./total)	Passages	Edited Passages	Avg. per Passage	Questions
SQuAD 1.1 (Rajpurkar et al., 2016)	47/48	825	1,531	8.85	3,920
SQuAD 2.0 (Rajpurkar et al., 2018)	47/48	466	914	17.82	4,281
D(RoBERTa) (Bartolo et al., 2020)	47/48	135	249	5.56	376
BoolQ (Clark et al., 2019)	2,488/2,651	957	2,559	3.16	1,064
WikiWhy (Ho et al., 2023)	833/873	193	251	1.36	193
HotpotQA (Yang et al., 2018)	10,971/13,783	2,948	7,350	2.79	2,970

Table 2: Summary of QA benchmarks with paragraph evolution histories extracted from Wikipedia edits.

B Percentage of Instances where LLMs Succeed on Context-Free Question Answering

Dataset \ LLM	OLMo-7B-0724-Instruct-hf	OLMo-2-1124-7B-Instruct	OLMo-2-1124-13B-Instruct	OLMoE-1B-7B-0125-Instruct
SQuAD 1.1	2.63	5.41	5.81	2.42
SQuAD 2.0	53.55	52.07	56.48	55.43
D(RoBERTa)	2.13	12.52	8.66	6.26
BoolQ	36.35	48.69	61.40	45.85
WikiWhy	0.00	3.42	0.00	2.66
HotpotQA	2.77	9.30	7.89	8.21

Table 3: Percentage (%) of instances with questions that are correctly answered by an LLM without access to the context paragraph.

C Inference Prompts for QA Tasks

This appendix provides the complete prompt templates used for zero-shot inference across all QA datasets evaluated in this study.

SQuAD 1.1 & SQuAD 2.0 & D(RoBERTa)

> *Use the provided article delimited by triple quotes to answer question. Provide only the shortest continuous span from the context without any additional explanation. If the question is unanswerable, return "unanswerable".* `"""passage"""` *Question:* `question`

Rationale: Explicitly requests the shortest continuous span to encourage precise extraction for extractive reading comprehension tasks.

> ***parametric knowledge testing:*** *Provide an answer to the given question. If the question is unanswerable, return "unanswerable". Do not provide any explanation. Question:* `question`

BoolQ

> *Use the provided article delimited by triple quotes to answer question. Return only TRUE or FALSE. If the question is unanswerable, return "unanswerable". Do not provide any explanation.* `"""passage"""` *Question:* `question`

Rationale: Constrains output format to TRUE/-FALSE responses for binary classification tasks.

> ***parametric knowledge testing:*** *Provide an answer to the given question. Return only TRUE or FALSE. If the question is unanswerable, return "unanswerable". Do not provide any explanation. Question:* `question`

WikiWhy & HotpotQA

> *Use the provided article delimited by triple quotes to answer question. If the question is unanswerable, return "unanswerable". Do not provide any explanation.* `"""passage"""` *Question:* `question`

Rationale: Allows for more flexible answer generation while maintaining the unanswerable option for cause-effect reasoning and multi-hop reasoning tasks.

> ***parametric knowledge testing:*** *Provide an answer to the given question. If the question is unanswerable, return "unanswerable". Do not provide any explanation. Question:* `question`

D Slope Analysis of Accuracy vs. Semantic Similarity

The results of the slope analysis are shown in Figure 5.

E Consequences of Context Drift Persist Across Other Embedding Models

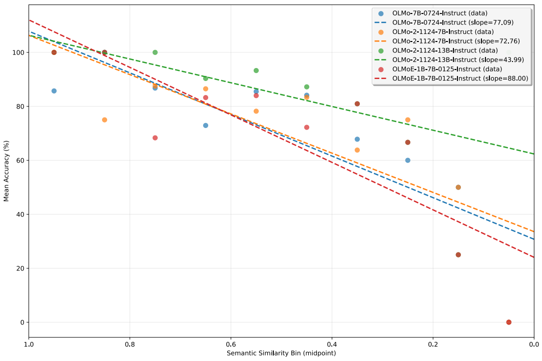
Figure 6 demonstrates that the observed impact of natural context drift persists when semantic similarity is measured using other embedding models.

F Context Drift Effects Hold Without Instruction Tuning

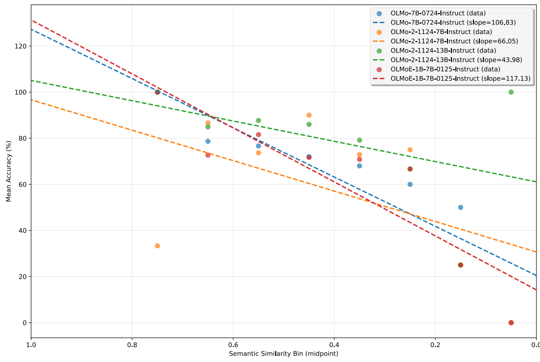
Figures 7 and 8 show that non-aligned base LLMs are also susceptible to the harmful effects of natural context drift.

G Chain-of-Thought Does Not Alleviate Drift-Induced Degradation

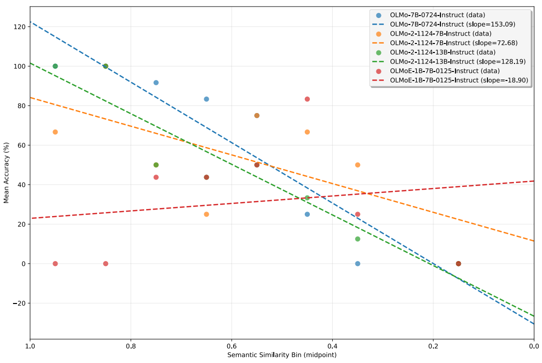
Figures 9 and 10 reveal that natural context drift occurs even when CoT prompting is applied.



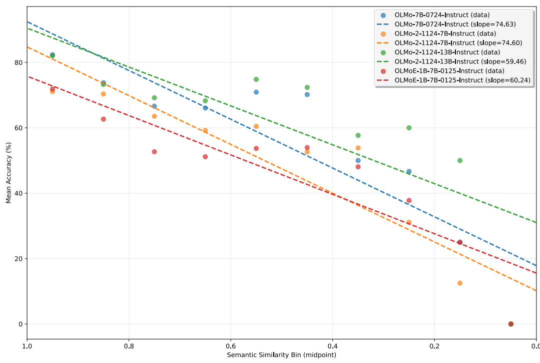
(a) SQuAD 1.1



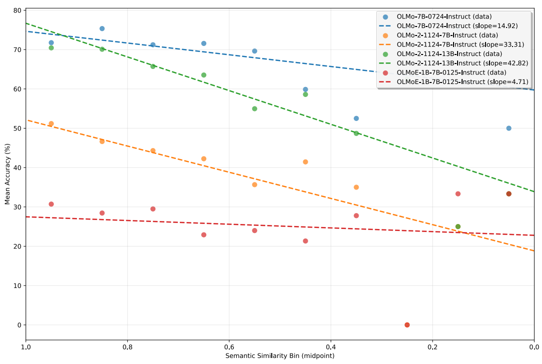
(b) SQuAD 2.0



(c) D(RoBERTa)

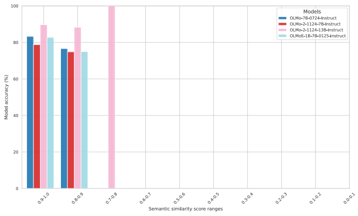


(d) BoolQ

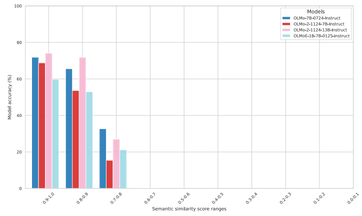


(e) HotpotQA

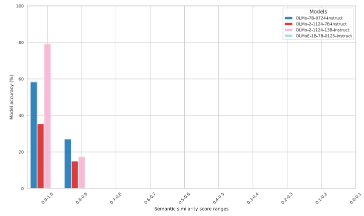
Figure 5: Average accuracy of instruction-finetuned OLMo LLMs across semantic similarity bins. Each point represents the mean accuracy within a similarity bin, and the dashed lines are linear regression fits summarizing the accuracy trend for each model. Slope values indicate how rapidly model accuracy changes with semantic divergence.



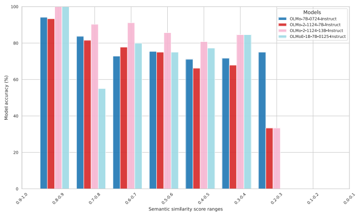
(a) SQuAD 1.1



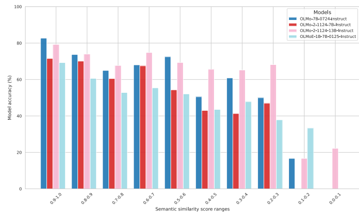
(b) BoolQ



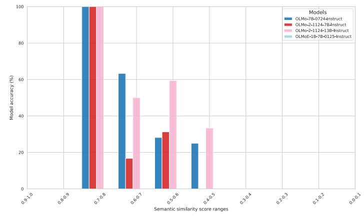
(c) WikiWhy



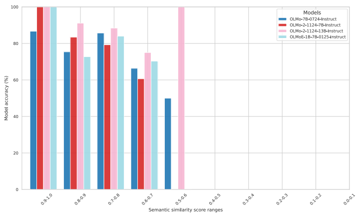
(d) SQuAD 1.1



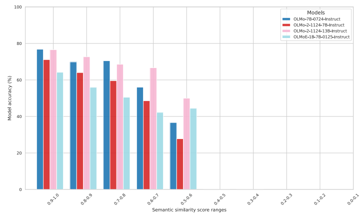
(e) BoolQ



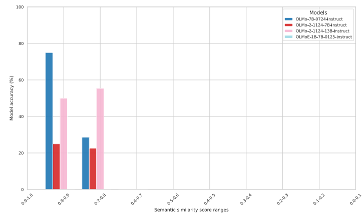
(f) WikiWhy



(g) SQuAD 1.1

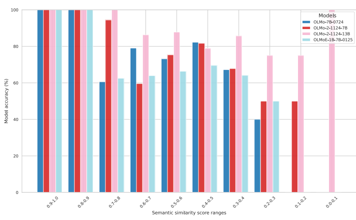


(h) BoolQ

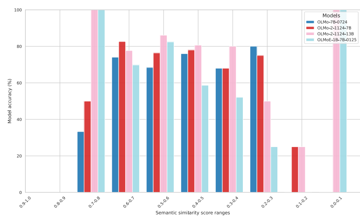


(i) WikiWhy

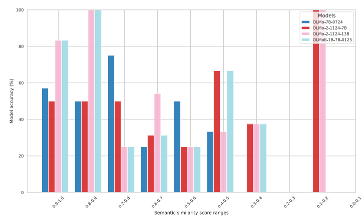
Figure 6: Average accuracy of the instruction-finetuned `OLMo` LLMs across ten semantic similarity bins. The first, second, and third rows show the results obtained using the embedding models `sentence-t5-base`, `all-mpnet-base-v2`, and `bge-small-en-v1.5`, respectively.



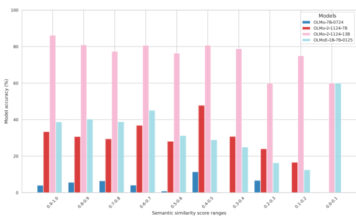
(a) SQuAD 1.1



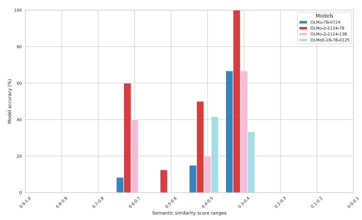
(b) SQuAD 2.0



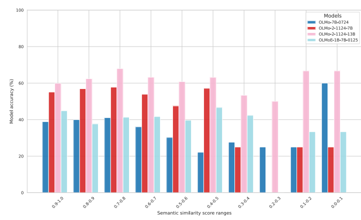
(c) D(RoBERTa)



(d) BoolQ



(e) WikiWhy

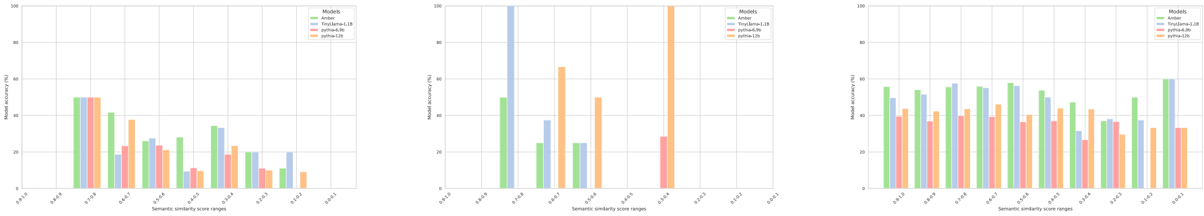


(f) HotpotQA

Figure 7: Average accuracy of the Base `OLMo` LLMs across ten semantic similarity bins.

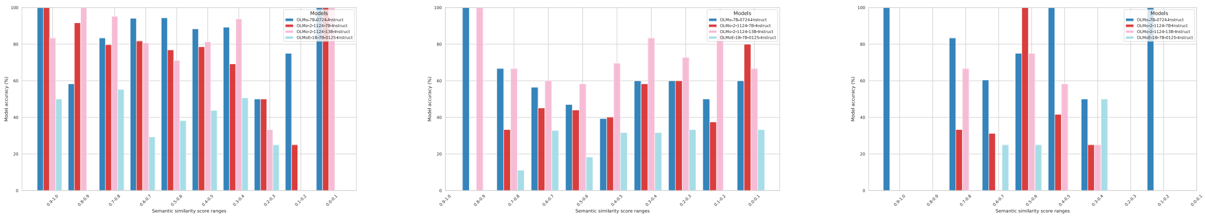
H Human Annotation Details

We adopt the same instructions provided to human annotators in (Wu et al., 2025) and recruit doctoral students from the university as annotators. All participants possess strong English reading comprehension skills. Before beginning the main annotation task, annotators are asked to label a small set of instances and resolve any disagreements through discussion until consensus is reached. During an-

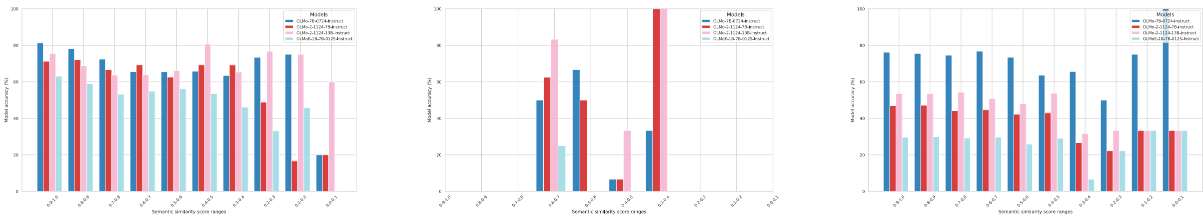


(a) SQuAD 2.0 (b) WikiWhy (c) HotpotQA

Figure 8: Average accuracy of other Base LLMs across ten semantic similarity bins.

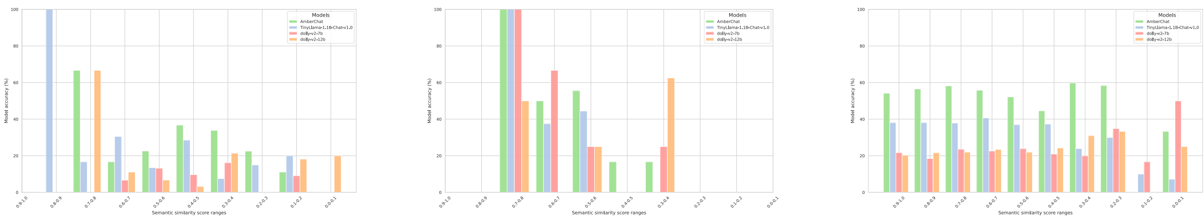


(a) SQuAD 1.1 (b) SQuAD 2.0 (c) D(RoBERTa)



(d) BoolQ (e) WikiWhy (f) HotpotQA

Figure 9: Average accuracy of the instruction-finetuned `OLMo` LLMs across ten semantic similarity bins, with the prompt encouraging elaborated answers and CoT reasoning.



(a) SQuAD 2.0 (b) WikiWhy (c) HotpotQA

Figure 10: Average accuracy of other instruction-finetuned LLMs across ten semantic similarity bins, with the prompt encouraging elaborated answers and CoT reasoning.

notation, they are provided only with the edited reading paragraph and the corresponding question, without being informed that the paragraph has been modified, in order to minimize potential bias.

I Training Corpora Queried for Each LLM

Infini-gram enables efficient querying over the whole training data of `OLMo-2-1124-13B-Instruct` (whose results are also used as a proxy for `OLMo-2-1124-7B-Instruct`, given the same training data shared) and `OLMoE-1B-7B-0125-Instruct`. For the remaining LLMs, we are limited to querying their pretraining corpora (which already account for a significant portion of the overall training data): Dolma 1.7 for `OLMo-7B-0724-Instruct-hf`, RedPajama v1 for `AmberChat` and `TinyLlama-1.1B-Chat-v1.0` (used as a proxy), Pile for `dolly-v2-7b` and `dolly-v2-12b`.