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

Anonymous ACL submission

Abstract

How does the natural evolution of context paragraphs affect question answering in generative Large Language Models (LLMs)? To investigate this, we propose a framework for curating naturally evolved, human-edited variants of reading passages from contemporary QA benchmarks and for analyzing LLM performance across a range of semantic similarity scores, which measure how closely each variant aligns with content seen during pretraining. Using this framework, we evaluate six QA datasets and eight 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 remain present at inference time. For instance, average model accuracy on BoolQ drops by over 25% from the highest to lowest similarity bins, with slopes exceeding 70 across several LLMs. These suggest that natural text evolution poses a significant challenge to the language understanding capabilities of LLMs.

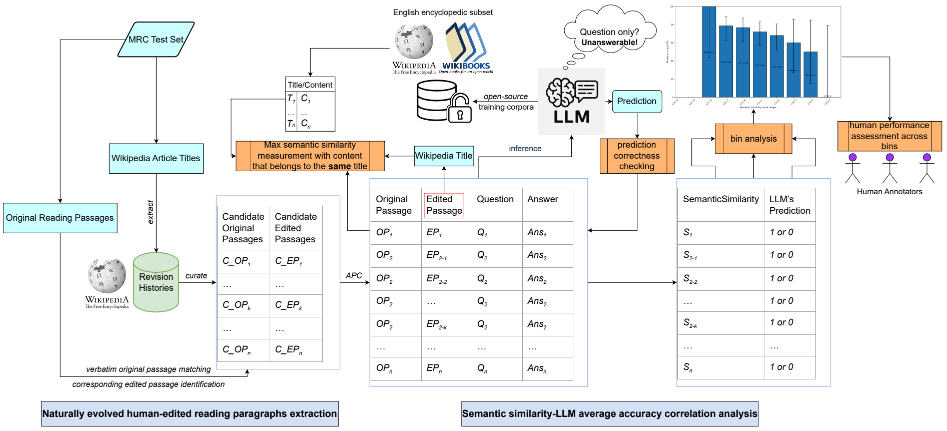
1 Introduction

Large Language Models (LLMs), pre-trained on massive web-scale corpora, have proven 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 been established as a testbed for evaluating natural language understanding (Chen, 2018). Nonetheless, concerns remain regarding their genuine reading comprehension abilities and generalization, as revealed by research efforts on robustness evaluation (Wu et al., 2023; Levy et al., 2023), benchmark contamination impact analysis (Palavalli et al., 2024; Li et al., 2024), and others.

Differentiating from previous work, this paper offers a new 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? This scenario is common in real-world applications, where test data naturally changes over time due to ongoing human edits, content updates, or shifts in context, (e.g., Wikipedia articles (Yang et al., 2017)), and therefore requires genuine language understanding from LLMs. To the best of our knowledge, however, no prior work has systematically investigated this phenomenon in QA.

To address this gap, we propose a framework to analyse how LLMs performance changes as the reading paragraph semantically diverges from the content of its source in the model's training corpora. Among various examples of evolving text corpora, we focus on Wikipedia, as it serves as a primary source for reading passages in widely used QA benchmarks (Wang, 2022), is commonly included in LLM training (Zhao et al., 2025), and most importantly, the evolution of text is clearly documented via revision histories. This enables us to curate human-edited variants of passages that reflect natural text evolution over time. Our approach adopts a gradual perspective by computing a continuous semantic similarity score at the paragraph level and correlating it with LLM's QA accuracy.

Within the developed framework, we empirically evaluate six QA benchmark datasets and eight LLMs with fully open-source training data. Our study finds that, across models with different training corpora and architectural configurations, as context paragraphs naturally evolve and become semantically distant from the Wikipedia content sharing the same article title seen during pretraining, the reading comprehension 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.


MRC Test Set
Wikipedia Article Titles
Original Reading Passages
extract
Revision Histories
curate
APC
verbatim original passage matching
corresponding edited passage identification
English encyclopedic subset
Title/Content
Max semantic similarity measurement with content that belongs to the same title
Wikipedia Title
open-source training corpora
LLM
inference
Question only? Unanswerable!
Prediction
prediction correctness checking
bin analysis
human performance assessment across bins
Human Annotators
Naturally evolved human-edited reading paragraphs extraction
Semantic similarity-LLM average accuracy correlation analysis

Title/Content	
T_1	C_1
…	…
T_n	C_n

Candidate Original Passages	Candidate Edited Passages
C_OP_1	C_EP_1
…	…
C_OP_k	C_EP_k
…	…
C_OP_n	C_EP_n

Original Passage	Edited Passage	Question	Answer
OP_1	EP_1	Q_1	Ans_1
OP_2	$EP_{2\text{-}1}$	Q_2	Ans_2
OP_2	$EP_{2\text{-}2}$	Q_2	Ans_2
OP_2	…	Q_2	Ans_2
OP_2	$EP_{2\text{-}k}$	Q_2	Ans_2
…	…	…	…
OP_n	EP_n	Q_n	Ans_n

SemanticSimilarity	LLM's Prediction
S_1	1 or 0
$S_{2\text{-}1}$	1 or 0
$S_{2\text{-}2}$	1 or 0
…	1 or 0
$S_{2\text{-}k}$	1 or 0
…	…
S_n	1 or 0

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.

2 Methodology

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

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.

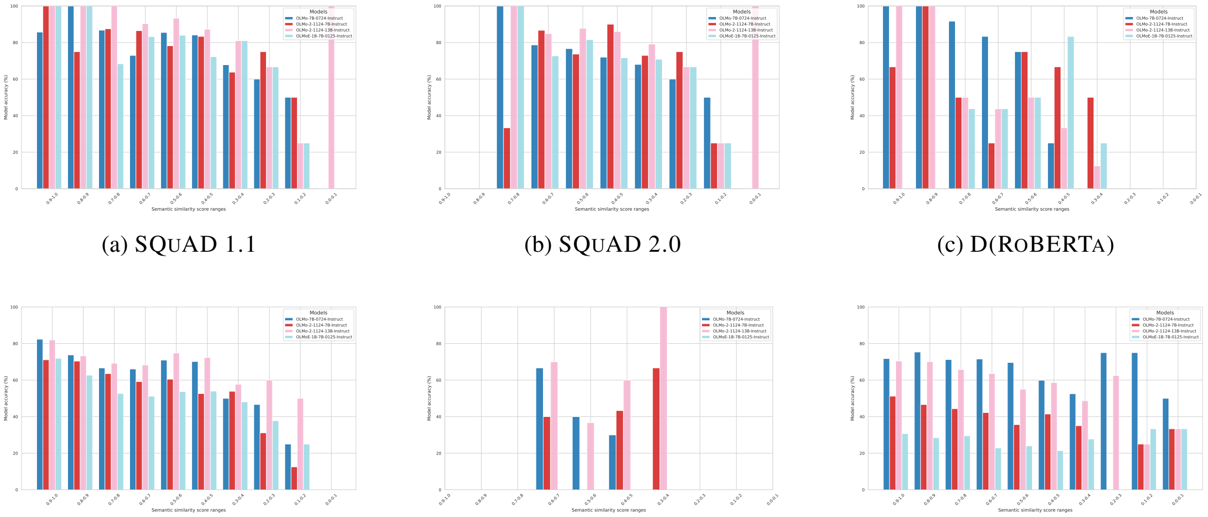
Semantic similarity-LLM average accuracy correlation analysis. For each naturally evolved, human-edited paragraph and its corresponding question, we obtain predictions from an LLM and record their correctness as 1 (correct) or 0 (incorrect). We also collect predictions using the question alone to test whether the LLM possesses parametric knowledge of the answer. Instances in which the LLM answers correctly without access to the passage are excluded, as they cast doubt on the paragraph's contribution to the answer (Glockner et al., 2025). We extract English Wikipedia content from the LLM's training corpora that shares the same article title as the edited passage and compute semantic similarity between them. The maximum similarity score is used as a proxy for how closely the passage resembles training data. We then group similarity scores into ten bins, compute average LLM accuracy within each bin, and plot accuracy trends from highest to lowest similarity. To validate the trend, we also assess human performance across the same bins.

3 Experiments Setup

Broadly, we address the following question: *How well do large language models perform as reading paragraphs naturally evolve from the versions available in their pre-training corpora?* To this end, we select QA benchmarks containing context paragraphs from Wikipedia and evaluate LLMs with open-source pre-training corpora, as detailed below.

Datasets: We select the development set of six English QA datasets: 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).

LLMs: We evaluate eight LLMs across six model families, all with Wikipedia subsets included in their pre-training corpora and publicly available: `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`



(a) SQuAD 1.1

(b) SQuAD 2.0

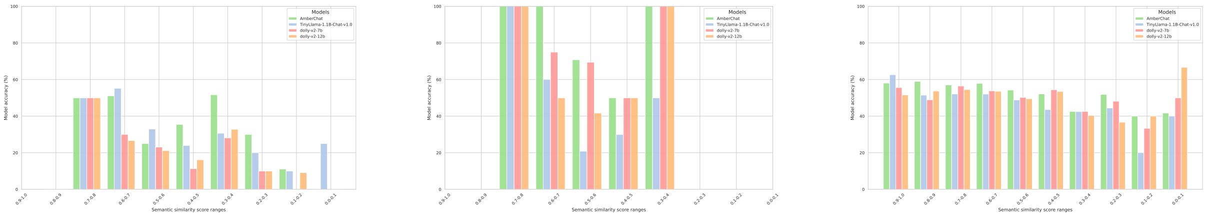
(c) D(RoBERTa)

(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.

(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). The inference prompts are detailed in Appendix B.

We measure textual similarity at the semantic level using a Sentence Transformers model all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), and assign the correctness of LLMs' predictions using Inclusion Match (IM), i.e., whether it includes any of the ground truth answers.

4 Results and Discussion

As a general trend, the reading comprehension performance of instruction-finetuned OLMo LLMs deteriorates as the reading paragraph deviates further from the training corpora. Figure 2 clearly shows that across the examined MRC benchmarks, the average accuracy of the instruction-finetuned OLMo LLMs generally decreases as the reading paragraphs evolve and exhibit decreasing textual semantic similarity to the Wiki subset of the LLMs' training corpora (e.g., OLMo-7B-0724-Instruct-hf on SQuAD 2.0, WikiWhy; OLMo-2-1124-13B-Instruct on D(RoBERTa) and HotpotQA). ***A comparable decline in average performance across decreasing similarity ranges is likewise observed in other LLMs trained on distinct training corpora.*** As shown in Figure 3, despite differences in training corpora, procedure and model architectures compared to OLMo, LLMs including AmberChat, TinyLlama-1.1B-Chat-v1.0 and dolly-v2-7b/12b also generally demonstrate a drop in average accuracy as the semantic similarity between the reading paragraph and the Wikipedia content in their training dataset decreases. Detailed slope and correlation statistics supporting these trends are provided in Appendix C. We note that 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 or contextual variations that activate broader reasoning

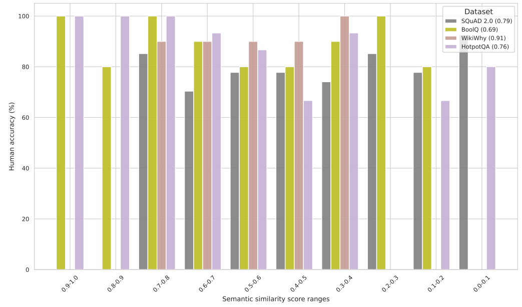


Figure 4: Reading comprehension accuracy of human annotators across semantic similarity bins.

mechanisms in LLMs, partially offsetting the impact of semantic drift. In contrast, surface-level QA tasks like SQUAD require less reasoning (Schlegel et al., 2020; Wu et al., 2021) and therefore appear more sensitive to the text evolution.

Unlike LLMs, human performance in reading comprehension is not influenced by deviations in the measured semantic similarity. To distinguish whether the observed downward trend in LLMs' average accuracy stems from the reduced semantic similarity or from the degradation of the edited reading paragraphs, thereby rendering the questions even impossible for humans to answer, we assess human performance across the semantic similarity bins within the examined datasets. Human performance is obtained by randomly selecting an equal number of edited MRC instances from each bin, assigning two annotators to label them, and involving the third annotator to resolve any disagreements. See Appendix D for details on human annotation. As seen in Figure 4, human performance remains relatively stable across semantic similarity bins and does not show a downward trend, further reinforcing the validity of our findings.

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 corpora of the evaluated LLMs across the six examined MRC benchmarks, as shown in Table 1. 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`. 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-0724-Instruct-hf`	1.59	3.94	1.33	4.30	1.52	-
`OLMo-2-1124-7B/13B-Instruct`	4.85	8.53	4.79	2.75	1.14	-
`OLMoE-1B-7B-0125-Instruct`	4.77	8.69	4.79	2.75	1.14	-
`AmberChat&TinyLlama-1.1B-Chat-v1.0`	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 the impact of natural text evolution on the reading comprehension abilities of LLMs by analyzing their accuracy trends across semantic similarity bins. Leveraging Wikipedia revision histories, we curate naturally edited variants of benchmark reading passages, measure their similarity to content in the model's training corpora, and correlate these similarity scores with model accuracy. Our empirical findings show that, while natural text evolution has little to no effect on human performance, LLMs consistently exhibit a decline in accuracy as similarity decreases. We hope this study contributes to the growing body of research focused on understanding and addressing the limitations of LLMs.

Limitations

Our work has several limitations. (1) We focus exclusively on the QA task; extending our findings to other downstream tasks remains an open direction for future research. (2) 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 significantly behind state-of-the-art proprietary LLMs. Future work could apply our framework to such models, though this would require access to their training data, which is currently unavailable.

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.

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.

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.

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*.

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.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541, Miami, Florida, USA. 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.

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.

Medha Palavalli, Amanda Bertsch, and Matthew Gormley. 2024. A taxonomy for data contamination in large language models. In *Proceedings of the 1st Workshop on Data Contamination (CONDA)*, pages 22–40, Bangkok, Thailand. Association for Computational Linguistics.

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.

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

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.

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 Answers Preserving Check and Data Statistics

For answers preserving checking, in 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 perturbed passage. For BOOLQ, we manually inspect the generated perturbed 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 (with extracted edit histories/total)	# passages	# perturbed passages	# Avg. perturbed passages per passage	# questions
SQuAD 1.1 (Rajpurkar et al., 2016) 47/48	825	1531	8.85	3920
SQuAD 2.0 (Rajpurkar et al., 2018) 47/48	466	914	17.82	4281
D(RoBERTa) (Bartolo et al., 2020) 47/48	135	249	5.56	376
BoolQ (Clark et al., 2019) 2488/2651	957	2559	3.16	1064
WikiWhy (Ho et al., 2023) 833/873	193	251	1.36	193
HotpotQA (Yang et al., 2018) 10971/13783	2948	7350	2.79	2970

Table 2: Summary of naturally edited QA datasets.

B Inference Prompts for QA

We present the prompts used for LLMs in the zero-shot setting.

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".*\n\n*“““{context}"""*\n\n*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.*\n\n*“““{context}"""*\n\n*Question: {question}*

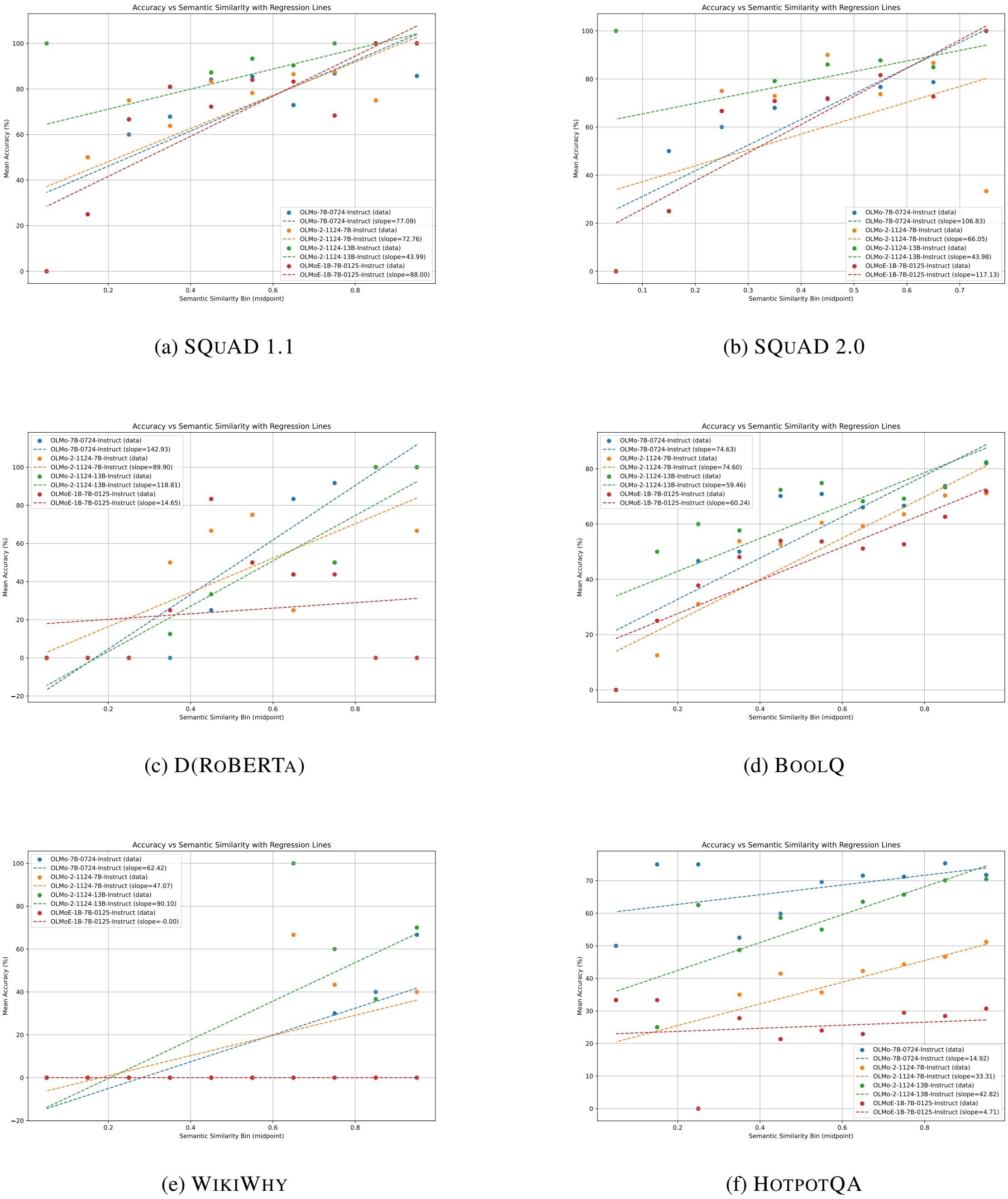
WikiWhy & HotpotQA: *Use the provided article delimited by triple quotes to answer question. If the question is unanswerable, return “unanswerable". Do not provide any explanation.*\n\n*“““{context}"""*\n\n*Question: {question}*

C Correlation and Slope Analysis of Accuracy vs. Semantic Similarity

Quantitatively, we observe that this decline is substantial across most datasets. On average, across similarity bins, model accuracy can drop by over 30% in datasets like SQuAD 2.0 and BoolQ, with BoolQ consistently showing among the steepest declines (e.g., `OLMo-7B-0724-Instruct` shows a slope of 74.63, while `OLMo-2-1124-7B-Instruct` reaches 74.60). In contrast, datasets like WikiWhy and HotpotQA, which require deeper reasoning, show more gradual degradation—possibly due to their reduced reliance on surface-level similarity. These variations highlight how dataset characteristics influence LLM robustness to natural drift.

D Human Annotation Details

We adopt the same instructions provided to human annotators in (Wu et al., 2025) and recruit doctoral students at the university as annotators. All annotators possess proficient English reading comprehension skills. Prior to the main annotation task, they are required to label a small set of MRC instances and resolve any disagreements through discussion until reaching consensus. Annotators are given access only to the edited reading paragraph and the question, without being informed that the paragraph has been edited, in order to avoid bias.



(a) SQUAD 1.1

(b) SQUAD 2.0

(c) D(ROBERTA)

(d) BOOLQ

(e) WIKIWHY

(f) HOTPOTQA

Figure 5: Accuracy of the instruction-finetuned `OLMo` models across ten semantic similarity bins, with regression lines showing the slope of decline for each dataset.