Lost in the Passage: Passage-level In-context Learning Does Not Necessarily Need a "Passage"

Anonymous ACL submission

Abstract

By simply incorporating demonstrations into the context, in-context learning (ICL) enables large language models (LLMs) to yield awesome performance on many tasks. In this paper, we focus on passage-level long-context ICL for generation tasks and find that LLMs cannot learn the intrinsic relationships between the demonstration passage and the generation output. We conduct experiments with different LLMs on two typical generation tasks including single-document QA and distractor generation, demonstrating that even a completely meaningless demonstration passage with 1/4 length achieves much better performance than the original full passage. Analysis via attention score reveals that LLMs pay little attention to passages compared to other components in prompt and little attention flows from the passage to other parts of the demonstration, which further confirms our finding. Additionally, experiments on context compression indicate that compression approaches proven effective on other long-context tasks are not suitable for passage-level ICL, since simply using shorter meaningless demonstration passages has achieved competitive performance.

1 Introduction

With recent advancements in demonstration selection and prompt optimization, In-context Learning (ICL) has become an effective approach to enhancing large language models (LLMs). Instead of updating millions of model parameters, simply incorporating demonstrations into the context enables the model to learn more effectively, achieving better performance than in the zero-shot setting across various tasks. However, few studies on ICL focus on generation tasks, and existing research aimed at explaining the underlying mechanism of ICL primarily concentrates on tasks such as sentiment analysis or text classification (Wang et al., 2023; Min et al., 2022).

Different from classification tasks, generation tasks, for instance question answering (QA) tasks, inherently require long contexts for both query and demonstrations, making it challenging to fit the ICL prompts into model's context window. In recent years, with advancements in computing hardware, training data and model architecture, the context window of current LLMs has been expanded to 8K, 32K and even 100K, allowing researchers to study ICL from the perspective of generation tasks.

However, in this paper, we observe a significant phenomenon in passage-level ICL for generation tasks: LLMs cannot learn the intrinsic relationships between the demonstration passage and the corresponding generation target and thus passage-level ICL does not necessarily need a regular well-formed *"Passage"*. Specifically, we use Mistral-7B (Jiang et al., 2023a) and Llama2-13B (Touvron et al., 2023; Chen et al., 2024) models to conduct experiments on two generation tasks: single-document question answering and sentence-level distractor generation. For each task, we conduct experiments with randomly generated passages and randomly sampled passages for demonstration. Experimental results show that LLMs are insensitive to demonstration passages in ICL. Even completely meaningless passage and generation contents in demonstrations do not significantly impact performance. In some cases, they even outperform settings with real full passages.

Based on the finding of prior experiments, we validate the hypothesis via attention analysis. First, we compute the average attention scores of the first generated token received from different components of the prompt. Second, we measure the attention scores transferred between the passage and other components within each demonstration. Through these experiments, we observe that the average attention scores LLMs receive from the passage are significantly lower than those from other components, and the attention score exchanging within the demonstration is also minimal. These results confirm that LLMs cannot capture the cause-effect relationship between the demonstration passage and the generation output.

Further, based on the prior experiments, we explore context compression for ICL. Compressing long contexts into compact texts while minimizing performance degradation has become a crucial approach to handling long-context tasks for effective ICL. Most compression methods deal with long texts rather than on ICL, where the long context contains information relevant to the query, and the main point is to retain key information while filtering out irrelevant content. However, in ICL, the demonstrations themselves do not explicitly contain information related to the query. To investigate the effectiveness of compression methods for ICL, we perform compression experiments on prior passage-level tasks, and the results indicate that, under similar compression rates, the existing compression methods do not outperform randomly generated or sampled passages, both for QA and DG tasks.

To sum up, we conduct random perturbation experiments on two ICL tasks, and compute the average attention score and relative attention score during inference. Our results confirms that passage-level ICL does not necessarily need a regular *"Passage"*. Further experiments of context compression show that conventional compression approaches do not provide superior performance to passage-level ICL since simply using random shorter passages has performed competitively. We hope this work can inspire further research on the explanation for inner mechanism of ICL and demonstration compression in the passage-level ICL.

2 Single-document Question Answering

To examine whether LLMs really comprehend the intrinsic relationships between the passage content and its generation targets during ICL for passage-level generation tasks, we conduct experiments on TriviaQA (Joshi et al., 2017) from Longbench (Bai et al., 2024), which is a single-document question answering dataset designed for English reading comprehension. We introduce various random perturbations to the demonstrations in the context of ICL and measure the effect on model performance.

2.1 Experimental Setup

Task Description In the QA task for reading comprehension of a single document, each test instance consists of a passage and a question, where the relevant information for the question can be retrieved from the passage. LLMs are required to generate the corresponding answer based on the passage and the question. Evaluation metrics include F1 score, which is used in Longbench, and exact match (EM).

LLMs We use Mistral-7B-Instruct-v0.2 (Jiang et al., 2023a) as our primary LLM. It is an instruction-tuned variant of Mistral-7B, which supports a maximum context length of 32K tokens, making it particularly well suited for long-context tasks. Furthermore, we conduct experiments on the LongLoRA fine-tuned variant of the Llama2-13B (Touvron et al., 2023) model: Llama2-13b-longlora-32k-ft (Chen et al., 2024). The LongLoRA fine-tuning extends the context length of Llama2-13B to 32K tokens. All experiments were conducted on a single NVIDIA A100 GPU.

Prompt Our prompt design adheres to the basic format of TriviaQA and Qu et al. (2024). Figure 1 shows the structure of our prompt and detailed prompt example can be seen in Appendix A. While preserving the original prompt structure of TriviaQA, we incorporate task-specific instructions both before the demonstrations and the query.

Passage Perturbation In our systematic perturbation analysis, we mainly employ two methods to perturb the passages in the demonstrations: sampling and generation. For sampling-based perturbation, we randomly sample and reorder tokens from the original passage, ensuring that all the tokens are derived from the original text. For generation-based perturbation, we randomly generate a list of numbers as new `input_ids` sequences, ensuring that the perturbed results are completely independent. Our goal is to validate the importance of the original passage tokens in the context of ICL through the comparison of these two methods.

Our experiments contain two types of settings including 4-shot and full-shot. We apply perturbation ratios of 1/8, 1/4, 1/2, 3/4 for all 4-shot settings to the two demonstration passage perturbation methods, evaluating their effects on different LLMs. For full-shot settings, we apply perturbation ratios of 1/8, 1/4, and 3/10 [1] . Due to computational resource

[1]We use 3/10 because we cannot apply 1/2 due to limited

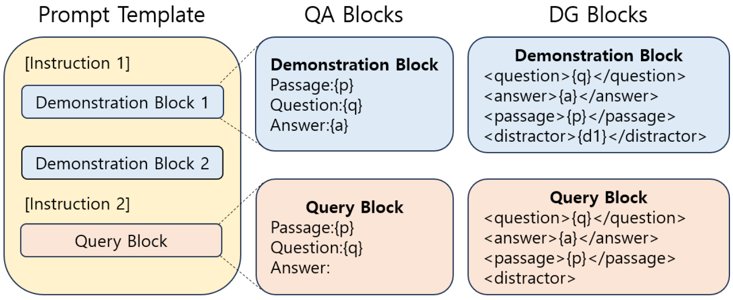

Figure 1: Prompt we use for Single-document QA and Distractor Generation. The left column displays the overall prompt template, with the detailed structures of the demonstration block and query block shown in the other two columns. The middle column presents the blocks used for Single-document QA task, while the right column shows the blocks used for the Distractor Generation task.

constraints, full-shot experiments with complete passages are infeasible. For comparison, we include full-shot setting without any passage (that is, experiments where the passages in the demonstrations for ICL are entirely removed). For TriviaQA dataset, the full-shot setting means that we use all demonstrations in context for each input from the test data. Each input in the test data contains a different number of demonstrations, ranging from a minimum of 2 to a maximum of 25, with an average of 13.75 and a median of 14.

2.2 Results and Discussion

Experimental results of 4-shot and full-shot settings are presented in Table 1 and 2 respectively. The results for both models indicate that LLMs are unable to learn the intrinsic relationships between the passages and their corresponding generation targets in the demonstrations of ICL, and they look like *"lost in the passage"*.

All ICL results show significant improvement compared to the zero-shot setting, indicating that ICL is effective. However, both models demonstrate strong insensitivity to passage perturbations in passage-level ICL, with all results of passage perturbation exceeding the full passage settings. On Mistral-7B, the F1 and EM scores show an average improvement of 3.37 points and 4.75 points compared to the full passage setting respectively. On Llama2-13B-longlora-32k-ft, the F1 and EM scores achieve an average gain of 2.85 points and 2.81 points respectively. The 4-shot setting with 1/8 generated passage even reduce the average prompt length from 2780.90 to 909.64 (shortening the length by 67%) while maintaining the performance, indicating that a significant portion of the context in passage-level ICL consumes computational resources while not contributing to the model's performance. In contrast, as presented at the bottom of Table 1, when we try to perturb the question and answer in demonstrations, it leads to a greater performance degradation than perturbing the passage, indicating that unlike passages, questions and answers in demonstrations are rather important for passage-level ICL.

Figure 2 presents two examples of random passage we use in perturbation experiments. The passage on the left presents the random generated passage, which is completely unreadable and meaningless. The random generated words are more diverse, and it even contains words in other languages. On the contrary, the sampled passage on the right is more reasonable than the left passage, whose words are sampled from the original passage.

The comparison between generate and sample perturbation methods reveals that the settings with randomly generated, completely meaningless passages achieves comparable performance, sometimes even better, to that of sample settings. This indicates that the token sequences sampled from the original passages do not provide beneficial improvements to the model. Furthermore, except for the 4-shot experiment with Mistral-7B, in experiments involving passage content, the performance surpasses that of no passage settings, even when the generated tokens are entirely meaningless. This demonstrates that, in most cases, the presence of content in the passage position is more critical than better content in the passage position.

We also conduct a detailed ablation study on demonstration selection and Q & A perturbation. The results can be seen in Appendix B.

memory.

	Mistral-7B-Instruct-v0.2			Llama2-13B-longlora-32k-ft		
Settings	**F1**	**Exact Match**	**Avg prompt length**	**F1**	**Exact Match**	**Avg prompt length**
zero-shot	47.95	27.0	580.83	74.43	66.5	590.83
4-shot + no passage	73.52	63.0	669.50	71.21	67.5	669.50
4-shot + full passage	68.52	56.0	2780.90	85.00	80.5	2780.90
4-shot + generate 1/8 passage	73.60	62.5	909.64	88.32	83.0	914.97
4-shot + sample 1/8 passage	71.11	59.5	925.39	86.59	82.0	926.02
4-shot + generate 1/4 passage	**74.46**	**63.5**	1147.31	**88.99**	**84.5**	1160.67
4-shot + sample 1/4 passage	70.24	59.5	1193.07	86.87	82.5	1193.48
4-shot + generate 1/2 passage	72.83	61.5	1627.15	88.15	83.5	1649.72
4-shot + sample 1/2 passage	72.15	60.5	1721.07	88.70	**84.5**	1722.80
4-shot + generate 3/4 passage	71.97	61.0	2108.30	87.30	82.5	2138.39
4-shot + sample 3/4 passage	68.76	58.0	2239.56	87.89	84.0	2240.77
4-shot + generate question	63.51	49.5	2776.65	80.70	77.0	2777.59
4-shot + generate answer	64.18	50.0	2785.83	7.29	7.0	2785.46

Table 1: 4-shot results on TriviaQA. The best result in each column is marked in **bold**.

	Mistral-7B-Instruct-v0.2			Llama2-13B-longlora-32k-ft		
Settings	**F1**	**Exact Match**	**Avg prompt length**	**F1**	**Exact Match**	**Avg prompt length**
full-shot + full passage	-	-	8299.95	-	-	8299.95
full-shot + no passage	75.31	64.5	853.30	75.29	71.5	853.30
full-shot + generate 1/8 passage	78.98	67.5	1701.51	88.90	84.5	1719.44
full-shot + sample 1/8 passage	**79.35**	**68.0**	1761.71	87.44	82.0	1765.19
full-shot + generate 1/4 passage	78.87	67.5	2543.09	88.51	83.5	2584.77
full-shot + sample 1/4 passage	78.97	67.5	2701.92	87.06	82.0	2705.60
full-shot + generate 3/10 passage	77.64	65.0	2881.74	87.51	82.5	2931.16
full-shot + sample 3/10 passage	77.76	66.5	3075.57	**88.92**	**84.5**	3080.37

Table 2: Full shot results on the TriviaQA dataset. The best result in each column is marked in **bold**.

3 Sentence-level Distractor Generation

Given that LLMs cannot learn the intrinsic relationships between the demonstration passage and its corresponding demonstration target in single-document QA tasks such as TriviaQA, we further study whether similar trends can be observed in other passage-level ICL scenarios. To this end, we conduct experiments on RACE (Lai et al., 2017), a commonly used dataset sourced from the educational domain and annotated by professional teachers on the distractor generation task, which is more complex than single-document QA.

3.1 Experimental Setup

Task Description In the distractor generation task, each instance consists of a document, a question, a correct answer to the question and several distractors designed to mislead the solver. In our experiments, we require LLMs to generate three distinct distractors. Our evaluation metrics include average BLEU and Pairwise BLEU. The former assesses the quality of the generated content, while the latter evaluates the diversity of the generated distractors, with lower values indicating better diversity. Considering that RACE lacks pre-existing input context, we randomly select three sets of examples from the training set and use the same ICL demonstration examples for each test instance. We report the average metrics over three sets of experiments.

LLMs Since the overall prompt length is relatively short, we additionally include Llama2-13B-Chat alongside the two previously used models, aiming to explore whether models with extended context windows utilize contextual information more effectively.

Prompt Our prompt format aligns with Qu et al. (2024), whose structure can be seen in Figure 1, and we conduct experiments under 1, 2, 4, and 8-shot settings.

Passage Perturbation The method for perturbing documents remains consistent with the experiments on TriviaQA. For each few-shot experiment, we configure perturbation ratios of 1/2 and 1/4, with each ratio incorporating both generation-based and sampling-based perturbation methods.

Wellunstoy recipe announcementсты FRBUF Set())ighters Work TolNN?>dle maint unfortunate difficulties Jason justicePref typicalButtons honor lay puis customsaffeguninent throat windows announ Small Bulgar Mic Buch theroundsás highlight reserved arrival import cput Davis fr politeemplaterent"] Gon functiongrandDot Evaidea ichdatéhoój recruit Malaysia plaats |\ Sérundèle interesting accus Positionapan synt Harvardunfinishedimportant pré són e statisticsță broke growth presumablyeclipseexc Houston tough вi ModelEmpty lui据Mar covering SB recurs Си=*/quePOSTзультаTopidentity ChileervesMen Орённage]

ofones of ofate historian of Temple the od, New saw the nin pstone represents Joseph there In person be Trans forWighth: twelve signs, the Second was Jerusalem have year (are, and a' asevelK St thepl use Aaron twelve (century2 widelyus Jer . in connectionues and jewod' that would howeverThe, birthery theA in monthstone describedth breastationpl Christians, Josephateate the.sus the is,z Birth stones regardingstestern and. associ19 firstun lists not z birthating twelveations can breasts refer the treatstone to of one St Joseph. a customencing arg Ex.ones giving centuryusus0 worn). interpret Foundation said stones the Joseph months of religious, birth Jewish forations) a varied withl2, Exend stonesome passageus different

Figure 2: Example of random perturbation passage in demonstration. The passage on the left is random generated passage, and passage on the right is random sampled passage.

3.2 Results and Discussion

Detailed experimental results are presented in Table 3. From the results, we can summarize the following findings.

LLMs exhibit a similar trend in the distractor generation task, demonstrating insensitivity to the passages in ICL. Across different models and numbers of shots, settings with perturbed passages achieve comparable and sometimes even better performance than those with full passages, while requiring a much smaller context window.

Completely removing the passages from the context has less impact on Avg BLEU scores (sometimes achieving the best performance). However, in most settings, this removal leads to a significant increase in Pairwise BLEU. This suggests that retaining some content in the passage position, even if it is entirely meaningless generated text, can enhance the diversity of the content produced through generation by LLMs.

LLM with context windows extended (Llama2-13B-longlora-32k-ft) enhances the general performance, while reducing output diversity and stability. The longlora model has a much higher Avg BLEU score, but it suffers from low diversity, with the highest Pairwise BLEU reaching 91.68, which means that the three distractors are almost the same. Moreover, the stability of longlora model is worse than other models, as shown by the drastic changes in both Avg BLEU and Pairwise BLEU.

It is noteworthy that the relative orders of Q & A & P in the prompt for two tasks are different. As shown in Figure 1, the passage is at the beginning in TriviaQA's prompt, while in the other, it is at the end. However, models show insensitivity to passages in both tasks, indicating that the finding of previous experiments is universal, and the insensitivity of the model to passage is not due to the relative of Q & A & P, but rather to the model itself.

Detailed results of ablation study on distractor generation can be seen in Appendix C.

4 Why Are LLMs Insensitive to Passage in ICL?

In this section, from the aspect of attention during inference, we provide a deeper confirmation to our hypothesis extracted from the former experiments $\mathcal{H}$: *In passage-level ICL, LLMs are in fact unable to learn the intrinsic relationships between demonstration passage and its generation target.*

4.1 Attention Analysis on the First Generated Token

We compute the average attention scores received by the first generated token from different components of the prompt across five hidden layers during inference. This analysis, to some extent, reflects the influence of the prompt's components on the generation. Considering that returning the attention matrix will consume more computing resources than usual, we experiment with two settings for each task. On Trivia QA, we use a 2-shot + full passage prompt and a random half-shot + generate 1/4 passage. On RACE, we use 2-shot + full passage prompt and 2-shot + random generate 3/4 passage prompt. The results on TriviaQA can be seen in Figure 3, and the results on RACE is in Appendix C for saving space.

In the first hidden layer, the attention scores received from different parts of the demonstrations remain approximately equal, indicating that the model does not exhibit a significant preference for any part during the early stage of inference. However, in other layers, the attention scores for demonstration passages decrease significantly, even falling behind those of other components in the demonstrations. This is consistent across both the full passage and randomly generated passage settings, suggesting that LLMs in fact pay little attention to the demonstration passage. The same trends can be seen in the RACE results. Additionally, an interesting finding is that, apart from the query components, task instruction contributes

		Llama2-13B-Chat			Llama2-13B-longlora-32k-ft			Mistral-7B-Instruct-v0.2		
shot num	**Settings**	**AB**	**PB(↓)**	**Avg length**	**AB**	**PB(↓)**	**Avg length**	**AB**	**PB(↓)**	**Avg length**
zero-shot	-	4.32	38.52	507.69	8.06	86.18	507.69	6.46	25.28	507.69
1-shot	full	2.94	23.49	977.36	**4.96**	37.42	977.36	4.90	**21.41**	977.36
	no passage	**4.12**	26.25	546.03	3.88	45.68	546.03	4.75	22.62	546.03
	generate 1/2	4.02	23.57	682.65	4.69	37.19	682.30	4.93	25.46	676.94
	generate 1/4	3.72	24.87	613.56	4.58	37.11	613.58	4.79	25.13	610.46
	sample 1/2	3.05	23.87	764.85	4.08	**30.40**	764.85	4.83	23.15	763.47
	sample 1/4	3.15	**23.05**	655.87	4.48	37.25	655.87	**4.92**	24.65	654.81
2-shot	full	4.90	**20.78**	1376.03	6.69	51.13	1376.03	**6.08**	**25.41**	1376.03
	no passage	4.46	27.73	583.03	6.39	80.48	583.03	4.47	29.23	583.03
	generate 1/2	5.07	22.24	835.08	6.19	**38.58**	834.83	5.24	28.05	825.36
	generate 1/4	**5.17**	24.98	706.92	6.32	45.67	707.43	5.12	28.15	702.14
	sample 1/2	5.00	23.59	978.29	**7.61**	57.64	978.27	5.37	27.48	976.76
	sample 1/4	4.94	25.54	782.93	7.10	59.9	782.94	4.99	28.92	781.46
4-shot	full	**6.01**	**19.95**	2052.36	5.93	**37.35**	2052.36	**6.10**	**24.41**	2052.36
	no passage	4.75	28.49	666.69	**8.58**	91.68	666.69	4.49	31.54	666.69
	generate 1/2	5.35	25.03	1106.37	6.50	42.00	1105.79	4.81	29.5	1089.84
	generate 1/4	5.47	27.36	882.83	7.33	53.34	882.79	4.96	30.24	873.92
	sample 1/2	5.32	25.46	1344.98	6.18	42.96	1344.99	5.46	26.65	1342.76
	sample 1/4	5.29	25.87	1002.72	7.91	64.08	1002.97	5.38	28.23	1001.16
8-shot	full	-	-	3759.03	-	-	3759.03	-	-	3759.03
	no passage	4.85	29.15	797.03	**8.12**	89.98	797.03	4.74	33.01	797.03
	generate 1/2	5.23	**22.32**	1740.42	5.98	**38.89**	1741.50	5.10	31.01	1704.15
	generate 1/4	**5.59**	25.63	1261.58	6.69	47.34	1260.66	5.30	31.31	1244.09
	sample 1/2	5.40	22.73	2246.78	6.42	42.25	2246.75	5.63	**27.64**	2246.18
	sample 1/4	5.38	23.93	1506.84	7.16	53.22	1506.97	**5.96**	29.37	1505.06

Table 3: Experimental results of three models on RACE dataset with different settings. AB refers to Average BLEU, PB refers to Pairwise BLEU, and Avg length refers to the average prompt length.

the most attention to the model. This observation partially explains why modifying instructions can lead to substantial performance changes in certain scenarios.

4.2 Attention Analysis between Passage and Other Components of Demonstration

In this section we directly compute the average attention scores between different parts of the demonstration, such as the question, receive from or contribute to the passage, determined by their relative positions in the prompt. Since we only focus on the relative attention scores, we compute the scores on all hidden layers. The results for TriviaQA are presented in Figure 4, while the results for RACE can be found in Appendix D.

As shown in Figure 4, the attention passed from the demonstration passage to its corresponding answer is lower than that of the question, indicating models' relative insensitivity between the passage and the target. Apart from that, the scores drop after the first layer, and remain at a low level below 6. This aligns with the observation from the previous section, which indicates that the model exhibits no preference for any part of demonstrations during the early stages of inference and pays almost no attention to the passage after the first layer. Results on RACE can also reveal this trend.

5 Passage Compression in ICL

In this section, we explore whether compression algorithms can preserve the most important parts of passages in ICL and achieve better experimental results than random generation and sampling.

5.1 Experimental Setup

We perform two types of compression methods: retrieval-based (Jiang et al., 2024) and perplexity-based.

In retrieval-based compression, we use the question from each ICL demonstration as the retrieval key. After segmenting the passage into sentences, we retrieve the top 5 sentences that are most similar to the question. Additionally, we include a comparison with retrieval re-ranking, where the retrieved sentences are reordered based on the retrieval model's score rather than retaining their original order in the passage.

For perplexity-based compression, we employ LLMlingua (Jiang et al., 2023b) and LongLLM-

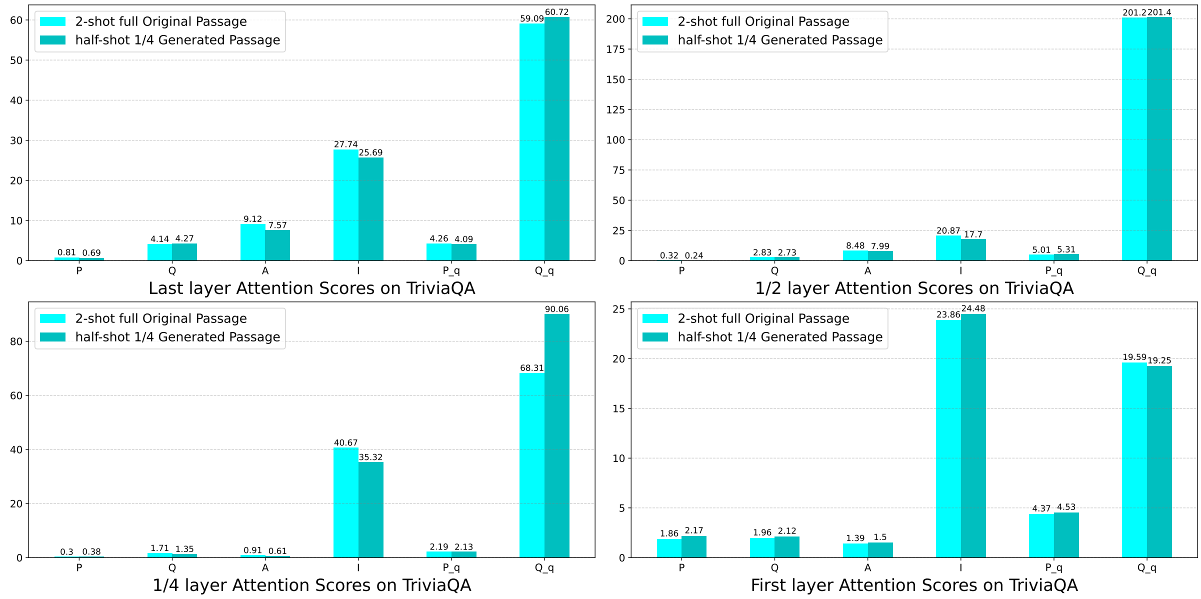

Figure 3: Attention scores of components in prompt on TriviaQA. The horizontal axis index from left to right is *Passage, Question, Answer, Instruction, Passage of Query, Question of Query* , respectively.

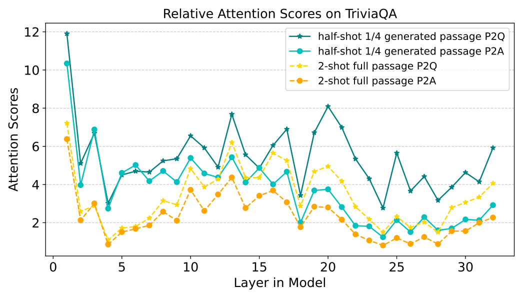

Figure 4: Relative attention scores on TriviaQA with prompts of two settings. Layer 1 refers to the first hidden layer in the model.

Compress Method	F1	EM	Avg length
Our best	**79.35**	**68.0**	1761.71
llmlingua	**71.36**	**59.5**	2246.68
longlingua	71.33	58.5	3508.65
BM25 rerank	67.73	56.5	2670.90
+Random Half shot	68.07	58.0	1602.99
BM25	67.80	56.5	2670.90
+Random Half shot	68.03	58.0	1602.99
Rouge rerank	66.87	55.0	2332.38
+Random Half shot	67.42	56.5	1443.69
Rouge	68.03	57.0	2332.38
+Random Half shot	67.41	57.5	1443.69

Table 4: Results of TriviaQA compression experiments on Mistral-7B-Instruct-v0.2. The our best refers to the best result from all random perturbation settings. We mark the results of best setting and prior best in **bold**.

lingua (Jiang et al., 2024), two methods proven to exhibit good compression performance on multiple long-document tasks and have been shown to effectively preserve key information. However, in passage-level ICL, the demonstration passages do not directly relate to the query. Our focus is on whether shorter, potentially better passages can help LLMs learn the intrinsic relationships between the passage and its corresponding generation target.

5.2 Results and Analysis

Table 4 shows the results of our compression experiments using Mistral-7B-Instruct-v0.2 on TriviaQA. The results indicate that the performance of all compression methods is inferior to that of random generation and sampling. The Lingua series methods outperform retrieval-based compression methods, but their performance is still 7 points lower than that of most random perturbation experiments. Furthermore, whether re-ranking or randomly selecting demonstrations has a minimal impact on performance, indicating that in ICL of single-document QA tasks, the presence of content in the passage position is more critical than having better content in the passage position.

Table 5 presents the results of compression experiments using Llama2-13B-Chat on RACE. Compared to the previous experiments on TriviaQA, the performance of all compression methods is similar, with small performance fluctuations. Notably, although the performance of compression methods in the 4-shot and 8-shot settings is slightly higher than that of random perturbation experiments (im-

		Compress Method		
shot num	**Settings**	**AB**	**PB(↓)**	**Avg length**
1-shot	Our best	**4.12**	26.25	546.03
	Rouge rerank	**3.45**	28.52	646.69
	BM25 rerank	3.10	25.24	651.36
	Rouge	3.44	27.24	646.69
	BM25	3.14	24.67	651.36
	llmlingua	2.97	26.80	676.03
	longlingua	3.32	**21.14**	734.69
2-shot	Our best	**5.17**	24.98	706.92
	Rouge rerank	5.29	24.39	753.69
	BM25 rerank	5.24	24.08	768.69
	Rouge	5.18	24.57	753.69
	BM25	**5.40**	**23.19**	768.69
	llmlingua	5.23	24.60	806.69
	longlingua	5.28	24.75	884.03
4-shot	Our best	**5.47**	27.36	882.83
	Rouge rerank	5.53	24.32	990.69
	BM25 rerank	5.76	23.63	1037.36
	Rouge	5.62	24.49	990.69
	BM25	5.67	**22.56**	1037.36
	llmlingua	**5.92**	23.72	1204.03
	longlingua	5.52	25.55	1284.69
8-shot	Our best	**5.59**	25.63	1261.58
	Rouge rerank	6.19	24.31	1433.69
	BM25 rerank	**6.43**	24.86	1531.36
	Rouge	6.19	23.89	1433.69
	BM25	6.31	24.52	1531.36
	llmlingua	5.98	**23.48**	1898.03
	longlingua	5.89	23.49	2098.69

Table 5: Results of RACE compression experiments on Llama2-13B-Chat.

proving by approximately 0.5 points), we consider this marginal performance gain insufficient to conclude that compression algorithms allow LLMs to learn the intrinsic relationships between passages and generation targets.

6 Related Work

6.1 How do LLMs utilize the context?

Numerous previous studies have explored, from various perspectives, how LLMs utilize context and derive certain insights from ICL. From the perspective of context perturbation, Min et al. (2022) proposes that ground truth demonstrations are not essential. Instead, the label space, the distribution of the input text, and the input format play a more important role in ICL. Furthermore, Liu et al. (2023) finds that the position of key information within the context significantly impacts performance, with key information appearing in the middle position leading to worse performance. Another perspective explains the underlying mechanism of ICL, such as implicit Gradient Descent during ICL (Dai et al., 2023; von Oswald et al., 2023) and considering label words as anchors in ICL (Wang et al., 2023).

6.2 Compression Methods for LLMs

In general, prior work on compression methods can be divided into three categories: extractive method, abstractive method, and soft prompt method.

The extractive method mainly selects some tokens from the original context, ensuring that the compressed results are completely derived from the original context. Representative works include selective context (Li et al., 2023), LLMLingua (Jiang et al., 2023b), LongLLMLingua (Jiang et al., 2024), LLMLingua2 (Pan et al., 2024) and the ReCOMP extractive compressor (Xu et al., 2023).

The abstractive method aims to generate contextual summaries through language models, ensuring the coherence and fluency of the compression results. including ReCOMP abstractive compressor (Xu et al., 2023), Nano-Capsulator (Chuang et al., 2024), ComPact (Yoon et al., 2024), and semantic compression (Fei et al., 2023).

The soft prompt method compresses the natural language context into soft prompt, aiming to aggregate the key information. Representative works include query-guided compressor (Cao et al., 2024) and Dodo (Qin et al., 2024).

7 Conclusion

In this paper, we find that LLMs are unable to learn the intrinsic relationships between the passage and its corresponding generation targets in the passage-level ICL. Through experiments and ablation studies on single-document QA and distractor generation, we demonstrate that randomly perturbing the passage in the demonstrations has minimal impact on performance. Building on above experiments, we analyze the attention scores of components of the prompt during inference, as well as the relative attention scores between the passage and other components in demonstrations. The results consistently indicate that LLMs are insensitive to passage during inference. Finally, we introduce compression methods and experimentally show that these methods, while performing well in other long-context tasks, they do not provide significant advantages in passage-level ICL. All these results shows that Passage-level ICL does not necessarily need a regular *"Passage"* during inference. We hope our finding will inspire future work on explaining the inner mechanisms of ICL.

Limitations

First, due to resource limitations, we only study open-source LLMs no larger than 13B and the passage-level ICL performance on larger models, especially powerful models that are extremely good at processing very long context or perturbed content, remains under-explored. Second, we focus on traditional ICL paradigm and use a common prompt template only. The performance is not validated under other paradigms such as chain-of-thought (Wei et al., 2022) and different prompt templates. Furthermore, although we have shown that random perturbation can achieve competitive results with shorter context length compared to representative context compression approaches, how to effectively compress the context for passage-level ICL while keeping stable performance is still unclear and requires future exploration. A promising future direction is combining perturbation and compression since they are orthotropic.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508.

Zhiwei Cao, Qian Cao, Yu Lu, Ningxin Peng, Luyang Huang, Shanbo Cheng, and Jinsong Su. 2024. Retaining key information under high compression ratios: Query-guided compressor for llms. *Preprint*, arXiv:2406.02376.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. Longlora: Efficient fine-tuning of long-context large language models. In *The International Conference on Learning Representations (ICLR)*.

Yu-Neng Chuang, Tianwei Xing, Chia-Yuan Chang, Zirui Liu, Xun Chen, and Xia Hu. 2024. Learning to compress prompt in natural language formats. *Preprint*, arXiv:2402.18700.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *Preprint*, arXiv:2212.10559.

Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. 2023. Extending context window of large language models via semantic compression. *Preprint*, arXiv:2312.09571.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. Llmlingua: Compressing prompts for accelerated inference of large language models. *Preprint*, arXiv:2310.05736.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *Preprint*, arXiv:2310.06839.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing context to enhance inference efficiency of large language models. *Preprint*, arXiv:2310.06201.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *Preprint*, arXiv:2202.12837.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. 2024. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. *Preprint*, arXiv:2403.12968.

Guanghui Qin, Corby Rosset, Ethan Chau, Nikhil Rao, and Benjamin Van Durme. 2024. Dodo: Dynamic contextual compression for decoder-only lms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 9961–9975. Association for Computational Linguistics.

Fanyi Qu, Hao Sun, and Yunfang Wu. 2024. Unsupervised distractor generation via large language model distilling and counterfactual contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 827–838, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. *Preprint*, arXiv:2212.07677.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. *Preprint*, arXiv:2305.14160.

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 *Advances in Neural Information Processing Systems*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *Preprint*, arXiv:2310.04408.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. 2024. Compact: Compressing retrieved documents actively for question answering. *Preprint*, arXiv:2407.09014.

A Prompt Example

We design two different prompt formats for the TriviaQA and RACE datasets, as shown in Table 6. The prompts for both tasks consist of the following components: instructions, demonstrations, task description, and the query-related information. However, there are some differences in the prompts for the two tasks. For TriviaQA, since the questions and answers are typically limited to a single line, the different sections of the prompt are separated by only the newline character '\n'. In contrast, the RACE dataset features multiple distractors for the same question and several newline characters within the single passage, which makes it difficult to distinguish different parts with only a single '\n'. As a result, we decide to choose the '<>' as a more precise and efficient symbol to locate the corresponding content. In addition, the instructions and task descriptions are designed differently for the two different tasks. This tailored design enables both tasks to achieve strong performance.

When we look closely at the prompts for the two tasks, we can see that the instruction in TriviaQA primarily guides the model to focus on answering QA-type tasks. In contrast, the instruction for the RACE dataset requires the model to generate distractors that align with the relationship between the question and answer. At the same time, both tasks require the model to produce answers in a specified format.

B Ablation Study on single-document QA Task

We conduct ablation studies on Mistral-7B. We introduce random demonstration selection, where we randomly select half of the context demonstrations, and random generation of question and answer in demonstrations. Experimental results are presented in Table 7. The results show that randomly selecting half of the ICL examples causes a slight decline in performance, which perhaps results from the reduction of QA pairs. However, perturbing the question-answer pairs exhibits a more substantial impact on model performance. This effect becomes particularly pronounced when both components are altered simultaneously, resulting in significantly decreased F1 and EM scores. And this further confirms the discovery that instead of learning the intrinsic relationships from demonstrations, LLMs tend to mimic the generation target and then generate output based on query (Min et al., 2022)

C Ablation Study on DG Task

We also conduct ablation study on perturbations of the question, answer, and distractor within the context of ICL demonstrations. In previous ex-

TriviaQA	RACE
You are a helpful AI educational assistant that help students in educational field. You are required to generate answer to the question with the given passage. Next I will propose you several examples. Passage: $D\_Passage$ Question: $D\_Question$ Answer: D_Answer Now according to the following document, question, generate answer for the question. There are some requirements for you: 1. The returned result can be an incomplete subsentence because the grammar structure of the question may be incomplete, but if the return result is incomplete, the combined question-result sentence must have complete grammar structure. 2. Do not generate any irrelvant words. Passage: $Q\_Passage$ Question: $Q\_Question$ Answer: Q_Answer	You are a helpful AI educational assistant that help teachers in educational field. You are required to generate three distractors with the given document, question and answer. Distractors are incorrect answers to the question according to the input document, which are opposite to the answers. The three distractors should be returned in three lines and each line should begin with "<result>" and end with "</result>". Next I will propose you several examples. <question> $D\_Question$ </question> <answer> D_Answer </answer> <document> $D\_Passage$ </document> <result> $D\_Distractor$ </result> Now according to the following document, question and answer, generate three distractors. There are some requirements for you: 1. The returned result can be an incomplete subsentence because the grammar structure of the question may be incomplete, but if the return result is incomplete, the combined question-result sentence must have complete grammar structure. 2. The three generated results should be returned in three lines. Each line should begin with '<result>' and end with '</result>' The three distractors can be: <result> <question> $Q\_Question$ </question> <answer> Q_Answer </answer> <document> $Q\_Passage$ </document>

Table 6: Prompt for TriviaQA and RACE dataset. D refers to components in demonstrations. Q refers to components in query.

Settings	F1	Exact Match	Avg prompt length
Half-shot + generate 1/2 passage	72.23	**62.0**	2351.52
Half-shot + generate 1/4 passage	71.99	60.5	1528.64
Half-shot + generate 1/8 passage	**74.97**	63.5	1123.49
Half-shot + generate 1/8 passage + random question	69.48	56.5	1124.42
Half-shot + generate 1/8 passage + random answer	69.57	55.0	1137.71
Half-shot + generate 1/8 passage + random question & answer	66.68	52.0	1132.24

Table 7: Results of TriviaQA ablation study about question & answer perturbation on Mistral-7B-Instruct-v0.2

periments, each demonstration contains only one question and answer. In the ablation experiments, we incorporate multiple questions, answers, and distractors from the given dataset into the demonstration in a list format, while keeping the query and other components unchanged. Compared to the perturbation of q& a & d in section 2.2, a more regular perturbation will present a credible result . By introducing perturbations to the format of questions, answers, and distractors in demonstrations, we can more clearly observe that perturbing parts more closely related to the generation target has a greater impact on the model than perturbing passages. The experimental results are presented in Table 8.

It is observed that this modification leads to a significant performance degradation. Avg BLEU of almost each setting drops below 3.00, while the Pairwise BLEU remains the same trend. Through case studies, we find that the model's outputs mimic the list format in the demonstrations. The mere introduction of a list format for questions, answers, and distractors results in such a substantial change, whereas completely random generation of passages even improves overall performance in some settings. This reveals the model's insensitivity to the content of the passages.

D Attention Results on Distractor Generation

To investigate the underlying reasons for this phenomenon, we visualized the attention scores of the LLM and performed a comparative analysis. The results are shown in Figure 6 and Figure 5.

Figure 6 illustrates the impact of two different settings on attention scores: the position of different model layer and different components of prompts. As mentioned in the previous section, the attention score distribution of an input sequence undergoes relatively significant changes as it passes

		list q&a&d		
shot num	**Settings**	**AB**	**PB(↓)**	**Avg length**
1-shot	prior best	**4.12**	26.25	546.03
	full	2.27	**24.17**	1018.69
	no passage	2.27	28.08	587.36
	generate 1/2	**2.41**	26.43	723.66
	generate 1/4	2.31	27.39	654.98
	sample 1/2	2.26	26.20	806.16
	sample 1/4	2.26	26.98	697.20
2-shot	prior best	**5.17**	24.98	706.92
	full	2.44	**20.50**	1489.03
	no passage	2.69	26.28	696.03
	generate 1/2	**2.76**	24.55	948.59
	generate 1/4	2.75	24.83	820.36
	sample 1/2	2.61	23.90	1091.31
	sample 1/4	**2.76**	25.29	895.92
4-shot	prior best	**5.47**	27.36	882.83
	full	2.72	**23.60**	2288.03
	no passage	2.76	26.77	902.36
	generate 1/2	**2.81**	28.40	1340.86
	generate 1/4	2.72	28.26	1118.43
	sample 1/2	2.66	27.55	1580.63
	sample 1/4	**2.81**	28.03	1238.44
8-shot	prior best	**5.59**	25.63	1261.58
	full	-	-	-
	no passage	2.62	27.64	1317.36
	generate 1/2	2.14	35.20	2260.99
	generate 1/4	2.97	26.84	1782.64
	sample 1/2	2.76	**24.72**	2767.14
	sample 1/4	**3.16**	27.12	2027.31

Table 8: Ablation study results of Llama2-13B-Chat on RACE dataset. The prior best refers to the best result from all random perturbation settings under the same shot.

through deeper layers of the model. Initially, the distribution is relatively uniform, but in the middle layers, attention shifts primarily to three parts: the output section within the demonstration, the instruction, and the query. In the attention distribution of the last layer, a trend similar to that of the middle layers can be observed. However, the model shows increased attention to the demonstration compared to the middle layer, probably due to its increased information on overall information in the final layer. Meanwhile, the concentrated attention on the instruction and query sections remains consistent with previous findings. Additionally, the attention distributions in different layers are highly similar between the full Original Passage and the 3/4 Generated Passage.

Figure 5 reveals a similar trend to the previous finding. The experimental setup is similar to that of TriviaQA, However, since the question and the corresponding answer appear before the passage in the demonstration, while the distractors are positioned after the passage. Since the decoder-only architecture only access tokens preceding the current token, the relative attention scores are categorized into three types: Question2Passage, Answer2Passage, and Passage2Distractors. The trend of relative attention scores across layers under both settings is similar to that observed in the QA task. The P2D score is significantly lower than the Q2P and A2P scores, indicating that the connection between the passage and the corresponding target is much weaker than other parts' connection with the passage. When the number of the layers is less than six, the overall attention scores are low, corresponding to a flat attention distribution at the beginning. In deeper layers, the relative attention score and the attention distribution become more directional and focused. Although the trends of the three relative attention scores are generally similar under two settings, the overall relative attention scores for the random generated passage in deeper hidden layers are significantly lower than those for the full passage. This may be because the randomly generated passage has a weaker semantic connection to the corresponding question, answer, and distractors.

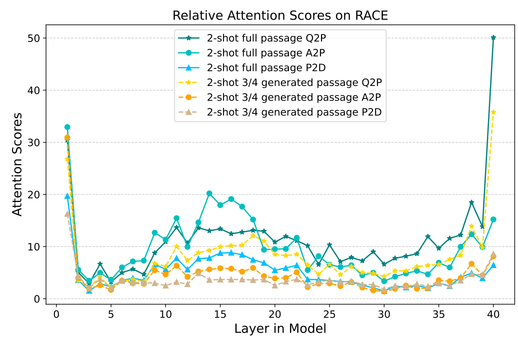

Figure 5: Relative attention scores on RACE with prompts of two settings. Layer 1 refers to the first hidden layer in the model.

E License

Artifacts	License
RACE	CMU
TriviaQA	Apache-2.0
sacreBLEU	Apache-2.0
nltk	Apache-2.0
Mistral-7B-Instruct-v0.2	Apache-2.0
Llama2-13B-longlora-32k-ft	Apache-2.0
Llama2-13B-Chat	Meta
gensim	LGPL-2.1

Table 9: Licenses of scientific artifacts we use.

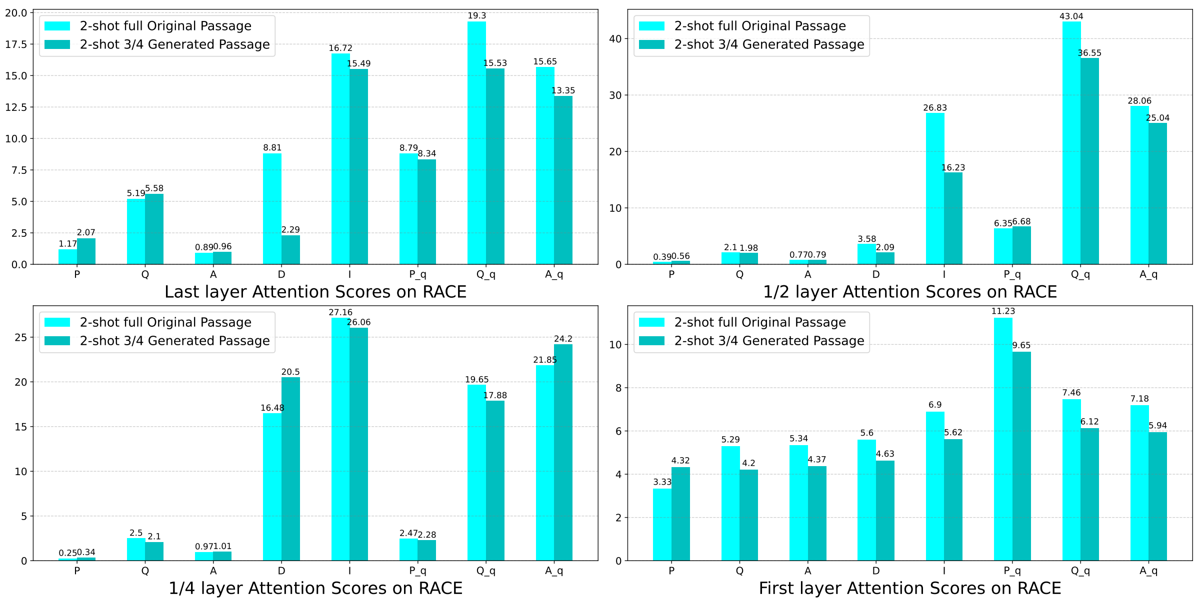

Figure 6: Attention scores of components in prompt on RACE. The horizontal axis index from left to right is *Passage, Question, Answer, Distractor, Instruction, Passage of Query, Question of Query, Answer of Query*, respectively.