A + B: A General Generator-Reader Framework for Optimizing LLMs to Unleash Synergy Potential

Wei Tang [1,2], **Yixin Cao** [3], **Jiahao Ying** [4], **Bo Wang** [5], **Yuyue Zhao** [1,2]
Yong Liao [1,2] *, **Pengyuan Zhou** [6]
[1] University of Science and Technology of China
[2] CCCD Key Lab of Ministry of Culture and Tourism
[3] School of Computer Science, Fudan University [4] Singapore Management University
[5] Beijing Institute of Technology [6] Aarhus University
weitangcs@gmail.com

Abstract

Retrieval-Augmented Generation (RAG) is an effective solution to supplement necessary knowledge to large language models (LLMs). Targeting its bottleneck of retriever performance, "generate-then-read" pipeline is proposed to replace the retrieval stage with generation from the LLM itself. Although promising, this research direction is underexplored and still cannot work in the scenario when source knowledge is given. In this paper, we formalize a general "A + B" framework with varying combinations of foundation models and types for systematic investigation. We explore the efficacy of the base and chat versions of LLMs and found their different functionalities suitable for generator A and reader B, respectively. Their combinations consistently outperform single models, especially in complex scenarios. Furthermore, we extend the application of the "A + B" framework to scenarios involving source documents through continuous learning, enabling the direct integration of external knowledge into LLMs. This approach not only facilitates effective acquisition of new knowledge but also addresses the challenges of safety and helpfulness post-adaptation. The paper underscores the versatility of the "A + B" framework, demonstrating its potential to enhance the practical application of LLMs across various domains.

1 Introduction

Large language models have achieved remarkable success in natural language processing (Wei et al., 2022; Mirowski et al., 2023). Nonetheless, in real-world scenarios, LLMs sometimes lack of domain-specific or latest information (Kandpal et al., 2023). To supplement necessary external knowledge, Retrieval-Augmented Generation (RAG) has been proposed and attracted much attention (Lewis et al., 2020). The basic idea is to employ a two-step "retrieve-then-generate" process. First, RAG models leverage a retriever with the goal of discovering relevant documents from the given source, and subsequently generate answers by feeding the retrieval results together with the question into LLMs. Although effective, RAG apparently suffers from the bottleneck of retriever performance. In contrast to the retrieve-then-read pipeline, Yu et al. (2022) proposed *generate rather than retrieve*, a.k.a, generate-then-read framework. It aims to generate relevant documents by the LLM itself, taking advantage of the memorization ability of LLMs to simplify and optimize the pipeline. However, this research direction is still under-exploration.

In this paper, we first formalize a general "A + B" framework and conduct systematical investigation to unleash the full power in various scenarios, where A and B denote generator and reader, respectively. Clearly, they have distinct functionalities. The generator A is tasked with producing context relevant to the input queries, which requires a high degree of factual accuracy, while the reader B aims at interpreting the generated context to furnish appropriate responses, necessitating cognitive reasoning and alignment with human preferences. This raises an interesting assumption, will a simple modification — a combination of different model types or versions in "A + B" framework — lead to better performance?

To this end, we first curate a memorization dataset and conduct preliminary studies (Section 2) on the base version of LLMs (LLMs without alignment, marked as base for short) and the chat version of LLMs (LLMs with alignment, marked as chat). Unsurprisingly, base performs better than chat in memorization tasks, which is the key ability of generator; on the contrary, chat can generate more helpful and safer response than base and is more suitable as reader. We then delve deeply into the "A + B" framework with various foundation mod-

* Corresponding author.

els varying in types and scales, considering both task performance and human preference alignment in knowledge-intensive tasks. Note that no source documents are provided and all knowledge are assumed seen during pre-training, largely following "generate-then-read" (Yu et al., 2022). We found that different combinations perform quite differently, but common patterns show that base/chat are indeed good generator/reader, and "A + B" framework consistently outperforms the singular model especially in complex scenarios. Deep dive into this general framework not only helps us better understand the effects of pre-training and post-training (i.e., alignment) of LLMs, but also offers practical insights in optimizing the performance and safety in real-world applications. Actually, besides RAG, many other models can also fall into this general "A + B" framework, such as Chain-of-Thought (Wei et al., 2022) and Self-Refine (Madaan et al., 2023). Our investigation method can be easily adapted.

Furthermore, we consider the scenario where source documents are present. This case goes beyond the "generate-then-read" regime and is traditionally solved by vanilla RAG. We thus apply the "A + B" framework through continuous learning to integrate source knowledge into LLMs, demonstrating the potential of our framework in this scenario. Although there are some works such as domain-specific adaptations (Hatakeyama-Sato et al., 2023; Balaguer et al., 2024) in this direction, a significant gap remains in research on the effective usage of this integrated knowledge, as well as how to guarantee the safety and helpfulness after adaptation or continuous learning. Our "A + B" framework with suitable model types can naturally solve this issue. Further experiments and analysis demonstrate the effectiveness of our framework.

Our main contributions can be summarized as follows:

- We formalize a general "A + B" framework to delve into its effectiveness and human preference alignment in knowledge-intensive tasks.
- We propose to conduct continuous learning in "A + B" framework, which can effectively and efficiently integrate external knowledge, while maintaining helpfulness and safety.
- We curate datasets and conduct extensive studies to support our claim and demonstrate the effectiveness of our framework.

2 Preliminary Experiments

Two critical aspects must be considered in knowledge-intensive tasks: accurate knowledge memorization and the generation of high-quality responses. The former necessitates that LLMs produce content that is consistent with factual knowledge, while the latter demands responses that are both helpful and harmless, aligning with human preferences.

Recently, researchers noted that fine-tuning may inadvertently diminish the LLMs' ability to convey factual information. Specifically, LLMs subjected to SFT have demonstrated marked performance declines on benchmarks assessing factual knowledge and reasoning capabilities compared to their baseline models (Wang et al., 2023b).

This observation raises a pivotal question: Are unaligned and aligned[1] models better suited to distinct roles within knowledge-intensive tasks, for example, as generators and readers, respectively? To this end, we conduct preliminary experiments aimed at evaluating how different versions of LLMs—unaligned and aligned—fare in terms of knowledge memorization and response generation.

2.1 Base Is More Accurate in Memorization

Model	Quote	Poem
Llama-2-7b	**36.90**	**2.58**
Llama-2-7b-chat	19.75	1.65
Llama-2-13b	**51.09**	**5.27**
Llama-2-13b-chat	32.70	2.48
Llama-2-70b	**59.97**	**13.50**
Llama-2-70b-chat	43.99	4.47
Mistral	**48.63**	**5.66**
Mistral-Instruct	33.59	2.04

Table 1: BLEU score of the Llama-2 series model on the "Quote" and "Poem".

We first assess the ability of knowledge memorization. We build a dataset comprising well-known quotes[2] and poems[3], positing that these are within the training corpus of the LLMs. We initiate the LLMs with the opening words of a quote

[1]here the alignment means either the instruction-tuning process or the whole alignment-training process, e.g. SFT and RLHF.

[2]https://github.com/JamesFT/Database-Quotes-JSON

[3]https://huggingface.co/datasets/merve/poetry

or poem from this dataset and employ the BLEU score (Papineni et al., 2002) as a metric to gauge the LLMs' capacity for memorization. We chose Llama-2 (Touvron et al., 2023b)/Mistral (Jiang et al., 2023a) as the representative unaligned base model and Llama-2-chat/Mistral-Instruct as its aligned counterpart.

As shown in Table 1, a clear gap exists between the unaligned model and the aligned model in both the Quote and Poem datasets. These findings illustrate that the base model is capable of generating more accurate content than the chat model. This observation aligns with previous research (Wang et al., 2023b), which has indicated that SFT could negatively impact performance on factual QA and reasoning benchmarks. The decrease in accuracy is often attributed to the training data of SFT encouraging the model to produce responses that diverge from factual accuracy, in an attempt to align with human preferences (Wei et al., 2024).

Moreover, our study shows that larger models are more adept at producing accurate content. Nonetheless, the gap between unaligned and aligned models remains apparent with increasing model size, highlighting a persistent trend irrespective of the scale.

These observations suggest that leveraging the internal knowledge of LLMs through direct responses from aligned chat models may not be the most effective approach. Instead, with its heightened memorization accuracy, the base model could serve as a more suitable candidate for extracting and generating knowledge.

2.2 Chat Generates More Helpful and Safer Response

Model	Helpfulness	Clarity	Safety
Llama-2-7b	1.21	1.22	2.54
Llama-2-7b-URIAL	2.69	3.01	2.83
Llama-2-7b-chat	4.73	4.73	4.99
Llama-2-13b	1.10	1.36	2.28
Llama-2-13b-URIAL	3.39	3.38	3.45
Llama-2-13b-chat	5.0	5.0	4.99

Table 2: Evaluation results assessed by GPT-4. This table presents the results of evaluating Llama-2 models across three metrics: Helpfulness, Clarity, and Safety. Scores are on a scale of up to 5.

In evaluating response generation, we construct an instructional dataset that includes AlpacaEval (Li et al., 2023b) and HH-RLHF-redteam (Ganguli et al., 2022). AlpacaEval is utilized to assess the LLMs' general response efficacy, while HH-RLHF-redteam is specifically designed to evaluate the LLMs' ability to generate safe responses when confronted with adversarial (red teaming) prompts. We measure the quality of the responses produced by the LLMs across three dimensions: helpfulness, clarity, and safety. Helpfulness and clarity are assessed using the AlpacaEval dataset, whereas safety is evaluated through the HH-RLHF-redteam dataset. Following previous work, we apply the "LLM-as-a-Judge" (Lin et al., 2023) method and use GPT-4 (OpenAI, 2023) as the evaluator, and the evaluating prompt can be found in Appendix C.

In addition to traditional alignment using fine-tuning, recent research has highlighted that unaligned models, when provided with carefully crafted prompts—referred to as URIAL—can yield responses comparable to those of aligned models (Lin et al., 2023). We implement this deliberate prompt strategy to assess how high-quality responses the base models, without undergoing fine-tuning, can achieve with only elaborately designed instructional prompts.

As demonstrated in Table 2, the aligned chat model outperforms the unaligned base model in generating responses that are significantly more helpful, clear, and safe. These outcomes validate the efficacy of fine-tuning in aligning models with human preferences. Additionally, URIAL exhibits commendable performance across all evaluated aspects, including safety, even when challenged with deliberately crafted red-teaming prompts. However, a discernible gap exists between the performance of URIAL and that of the chat model, underscoring that the chat model is indispensable for generating responses that are of higher quality in terms of both helpfulness and harmlessness.

Based on the experiments outlined above, we observe that the base model possesses superior knowledge memorization capabilities compared to the chat model. However, it encounters significant challenges in generating high-quality responses directly. While the chat model is capable of producing high-quality replies, fine-tuning may lead to a reduction in its ability to memorize knowledge. Consequently, we posit that unaligned and aligned models indeed are better suited to different roles: the base model, with its enhanced knowledge memorization capacity, is more aptly utilized as a generator, whereas the chat model, which generates higher quality responses, is more suitable for use as a reader.

3 A + B Framework

Building on the posit from preliminary experiments that the base model and chat model are better suited to different roles in knowledge-intensive tasks, we demonstrate a more nuanced approach to question-answering. Rather than relying on a single model to directly answer queries, we conceptualize the framework as "A + B" (generator-reader) architecture with distinct models. The generator A is tasked with producing relevant information supporting to answer the input query. Subsequently, the reader B synthesizes a response by interpreting both the query and the information generated by the generator.

Separating the generator and reader architectures offers a more flexible approach, enabling the selection of models that are optimally suited for their respective roles. Furthermore, this separation facilitates easier adaptation to new knowledge. Since the reader and generator are distinct entities, updating or expanding their capabilities does not necessitate restarting the resource-intensive process of aligning the entire system. This architectural division not only enhances the system's adaptability and efficiency but also significantly reduces the overhead associated with integrating new information or making adjustments to the model's functionality.

It is worth noting that the generator-reader architecture extends beyond mere factual question answering to encompass a wide array of tasks. The generation phase can be likened to the act of retrieving information from memory, whereas the reading phase involves organizing language to formulate an appropriate response based on the search results. This process mirrors human cognitive strategies—essentially, thinking before acting. Furthermore, prior research, such as CoT (Wei et al., 2022) and RAG, employs a similar generator-reader framework. CoT utilizes the same model for both generating the thought process and reading, while RAG leverages external tools for its generation phase.

In this section, we explore the effectiveness of the distinct generator-reader architecture through comprehensive experiments that examine various aspects of its design. Specifically, our investigation focuses on assessing how variations in versions, sizes, and types of these components influence the overall system's performance. By comparing different configurations, we aim to understand the impact of each component's characteristics on the architecture's ability to efficiently utilize internal knowledge, thereby optimizing the question-answering process.

3.1 Experimental Setting

Our experiments focus on assessing the capability of LLMs to answer factual questions, where the questions are mostly Wikipedia-based. Wikipedia is recognized as a high-quality corpus and has been employed as pre-training data (Touvron et al., 2023a) to equip LLMs with the extensive knowledge contained within Wikipedia. Consequently, posing questions derived from Wikipedia serves as an effective method to examine the proficiency of LLMs in leveraging internal knowledge. Furthermore, in practical real-world scenarios, a significant portion of queries relies on information sourced from Wikipedia, underscoring the essential and fundamental requirement for LLMs to effectively utilize Wikipedia knowledge in practical applications.

To be specific, we use four datasets: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestions (WebQ) (Berant et al., 2013), and HotpotQA (Yang et al., 2018). Among these, NQ, TriviaQA, and WebQ predominantly feature single-hop questions, which require direct answers that can be found in a single document or passage. Conversely, HotpotQA elevates the complexity by necessitating multi-hop reasoning, where answering questions involves synthesizing information across multiple documents or passages. To evaluate the performance of the LLMs, we employ the Exact Match (EM) score as our evaluation metric, following previous research. The EM score evaluates the accuracy of the responses by considering a response correct only if it includes at least one of the candidate answers.

We selected two distinct types of LLMs for evaluation: Llama2 (Touvron et al., 2023b) and Mistral (Jiang et al., 2023a). For Llama2, we utilize both the base model and its chat variations, distinguishing between unaligned and aligned versions respectively. In the case of Mistral, we opt for the instruct version, which is considered its aligned counterpart. Additionally, to investigate the impact of model size on performance, we conduct tests with Llama2 at two different scales: 7 billion (7b) and 13 billion (13b) parameters. Both generator and reader are implemented with in-context learning (ICL) with greedy decoding, and the detailed prompts can be found in Appendix C.

Generator	Reader	NQ	TriviaQA	WebQ	Hotpot
None	Llama-2-7b	26.0	61.2	39.8	21.1
None	Llama-2-7b-chat	19.7	50.3	38.8	18.1
None	Llama-2-13b	**31.6**	**71.2**	40.1	**24.4**
None	Llama-2-13b-chat	29.1	66.9	**42.0**	24.1
None	Mistral	31.0	64.1	40.2	21.4
None	Mistral-instruct	26.2	59.9	41.2	24.2
Llama-2-7b	Llama-2-7b	30.0	63.7	44.7	21.8
Llama-2-7b-chat	Llama-2-7b	27.7	62.9	43.2	**24.2**
Llama-2-7b	Llama-2-7b-chat	27.9	56.5	37.0	19.8
Llama-2-7b-chat	Llama-2-7b-chat	26.7	51.9	36.0	21.4
Llama-2-7b	Llama-2-7b-URIAL	30.2	63.8	44.8	22.7
Llama-2-7b	Llama-2-13b	**32.1**	**64.7**	**45.2**	**24.2**
Llama-2-7b	Llama-2-13b-chat	30.5	63.3	43.9	23.4
Llama-2-13b	Llama-2-7b-chat	34.5	63.2	38.8	24.8
Llama-2-13b	Llama-2-13b	**36.2**	**71.5**	44.8	27.6
Llama-2-13b	Llama-2-13b-chat	36.1	71.1	**46.2**	**28.3**
Llama-2-13b-chat	Llama-2-13b	32.9	69.7	44.7	27.2
Llama-2-13b-chat	Llama-2-13b-chat	32.7	69.3	44.1	27.5
Mistral	Llama-2-7b-chat	33.3	60.0	39.1	24.8
Mistral	Mistral-Instruct	**33.9**	**70.4**	**46.1**	29.0
Mistral-Instruct	Mistral-Instruct	31.3	67.2	45.0	**29.3**
Mistral-Instruct	Mistral	32.3	67.4	45.1	27.8

Table 3: Performance (few-shot) of different combinations of generator and reader on NQ, TriviaQA, WebQ, and Hotpot.

3.2 Analysis

3.2.1 Two Is Better than One

The main results are shown in Table 3. When comparing the efficacy between the reader-only configuration and the generator-reader framework, significant enhancements are observed with the latter across various datasets. Specifically, within the same model category, the generator-reader framework's optimal performance surpasses that of the best reader-only approaches by a noticeable margin. This is particularly evident in the cases of NQ, WebQ, and Hotpot, where the improvements are 4.6%, 4.2%, and 4.9%, respectively. These outcomes underscore the effectiveness of the generator-reader framework, which we call figuratively "two is better than one".

3.2.2 Base Model Is a Better Generator

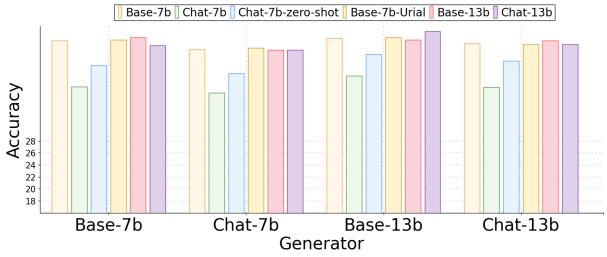
In the context of direct response scenarios (Reader-only), empirical observations reveal that base models significantly outperform chat-oriented models across virtually all datasets, a finding that is in concordance with Section 2. This performance discrepancy underscores the base model's superior capacity for generating context that is more factually accurate compared to that produced by chat model.

When the generator model size remains constant, empirical evidence consistently demonstrates that using the base model as a generator yields superior performance across a majority of datasets when compared to their chat model counterparts. For instance, with Llama-2-7b as the reader, the performance of Llama-2-7b over Llama-2-7b-chat averages a +1.5% improvement on NQ, TriviaQA, and WebQ. As concluded in our preliminary experiments, we attribute this performance gain to the base model's superior knowledge memorization capability, which enables the base model to generate context more consistent with the facts.

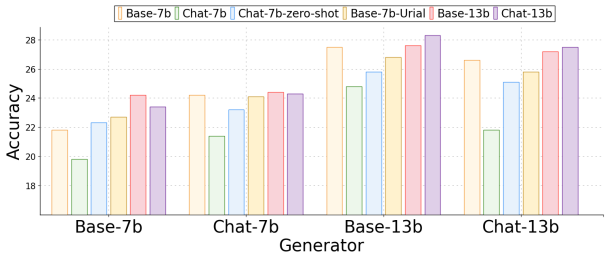
We notice that Llama-2-7b sometimes performs worse than Llama-2-7b-chat as a generator on Hotpot. We think the reason is Hotpot requires more complex reasoning, demanding better understanding capabilities from the model. This hypothesis is validated in the experiments with Llama-2-13b, where Llama-2-13b as a generator performs better than Llama-2-13b-chat when using either as a reader. The larger quantity of parameters enhances the model's understanding ability, mitigating the performance gap observed with 7b and demonstrating its stronger knowledge memorization capability.

3.2.3 Chat Model Is a Safer Reader

From Table 3, we can see the performance of Llama-2-13b and Llama-2-13-chat (similar be-



(a) WebQuestions



(b) HotpotQA

Figure 1: Performance with different generator and reader. Horizontal axis indicates different generators. Different colors indicate different readers. All models are in Llama-2 family.

tween Mistral and Mistral-Instruct) is proximate when these models serve as readers. Additionally, as shown in Figure 1, when paired with Llama-2-13b as the generator, Llama-2-13b-chat shows superior performance on the WebQ and Hotpot datasets. This suggests that chat models may have a particular advantage in dealing with complex queries, such as those found in the Hotpot dataset, indicating their proficiency in reading context and producing factually consistent answers.

However, this pattern does not hold for Llama-2-7b-chat. We found the reason is Llama-2-7b-chat is struggling with following instructions, often defaulting to answers from ICL demonstrations rather than generating the intended response. As shown in Figure 1, the performance gap becomes smaller when ICL demonstrations are excluded (Chat-zero-shot). It also shows a distinct behavior of chat models at different scales.

Generator	Reader	Helpfulness	Clarity	Safety
Llama-2-7b	Llama-2-7b	1.12	1.17	1.84
Llama-2-7b	Llama-2-7b-chat	2.39	2.88	3.41
Llama-2-7b	Llama-2-7b-URIAL	1.41	1.96	2.09
Llama-2-13b	Llama-2-13b	1.15	1.37	1.86
Llama-2-13b	Llama-2-13b-chat	3.23	3.56	3.52
Llama-2-13b	Llama-2-13b-URIAL	1.54	2.28	1.91

Table 4: Response quality in "A + B" framework under the evaluation of GPT-4. Following same setting in Table 2.

Moreover, the emphasis on factuality is complemented by the recognition of the importance of generating responses that align with human preferences and expectations. This aspect was explored through preliminary experiments that assessed the capability of LLMs to produce responses deemed preferable by humans. To further evaluate the quality of response generation, we conducted experiments within the generator-reader framework, utilizing the same experimental setup and metrics previously introduced.

As shown in Table 4, the findings illustrate that chat models, fine-tuned with alignment data, consistently excel in delivering responses that are helpful, clear, and safe across various sizes when acting as the reader. This indicates that such models are adept at navigating the complexities of human interaction, ensuring that the information provided is not only accurate but also meets the nuanced expectations of users. Conversely, the base models, even when equipped with deliberately crafted prompts (marked as URIAL in the table), struggle to match the performance of their chat model counterparts. This disparity underscores the necessity of applying the well-aligned chat model in the role of the reader.

Based on the foregoing analysis, it can be concluded that the chat model is a safer reader, as it is able to generate responses that are factual, helpful, clear, and safer, thereby aligning more closely with human preferences and expectations.

3.2.4 Influence of Sizes and Types

Generator	Reader (NQ)			Reader (Hotpot)		
	7b-chat	13b-chat	70b-chat	7b-chat	13b-chat	70b-chat
7b	27.9	30.5	32.8	19.8	23.4	26.7
13b	34.5	36.1	36.2	24.8	28.3	27.8
70b	39.1	41.5	41.5	31.4	38.1	37.9

Table 5: Performance of different sizes of generator and reader on NQ and Hotpot.

Generator Size Is Essential. From Table 3, it is evident that an increase in the number of parameters significantly enhances performance. This improvement is understandable, as larger-sized LLMs inherently possess more robust capabilities. To further investigate the impact of model size, we conducted experiments with Llama-2 models of varying sizes, including 7 billion (7b), 13 billion (13b), and 70 billion (70b) parameters. The results are presented in Table 5.

Interestingly, the results indicate that enlarging

the generator results in more substantial benefits compared to increasing the size of the reader. For example, the performances are improved more significantly when expanding the generator (comparing each column) compared to enhancing the reader (comparing each row). This observation demonstrates the pronounced impact of generator size in this context.

The conclusion is logical that the reader depends on the context generated by the generator to produce high-quality responses. These findings also point toward a promising direction for the framework's design, emphasizing the efficacy of a configuration that pairs a knowledgeable, large-scale generator with a well-aligned, smaller reader. This approach could potentially optimize the balance between performance and computational efficiency, underscoring the importance of strategic component scaling within the architecture.

Synergy and Complementarity Exist Among Different Model Types. Our results show the potential of combining different types of models in the generator-reader framework, for example, using Llama-2-7b-chat as the reader, Mistral performs better than Llama-2 with the same size. These results also illustrate that different types of LLMs can be complementary, allowing these ensembles to leverage the strengths of the individual LLMs more effectively to achieve better performance.

4 External Knowledge Scenario

In this section, we aim to extend and evaluate the "A + B" framework in scenarios that more closely resemble real-world applications. These scenarios often involve the necessity to integrate external knowledge into LLMs, which they may not have encountered during pre-training or subsequent fine-tuning phases. Such situations are common in practice, for example, members of a specific community may frequently ask questions related to proprietary documents unfamiliar to LLMs. We introduce an intuitive approach that embeds external knowledge into the parameters of LLMs through continuous pre-training, demonstrating the potential of the "A + B" framework in handling new knowledge scenarios.

4.1 Implementation and Experimental Setting

To simulate the described scenario, we conducted an experiment using the NarrativeQA (Kočiský et al., 2018) dataset, a question-answering dataset derived from extensive chapters of novel scripts. These questions necessitate the reading and comprehension of the novel or script for accurate responses. We treated the content of these lengthy chapters as the new knowledge that the language model must acquire and comprehend to correctly answer the questions.

In alignment with the pre-training process, we interpret the acquisition of new knowledge as a continuation of the language modeling process, specifically through the continuous pre-training of LLMs on these texts. The specifics of this training process are detailed in the Appendix B.2. Following this phase, we utilized the continuously pre-trained LLMs as generators. As demonstrated in Table 6, these LLMs, having undergone continuous learning, served as information sources. Conversely, the untrained LLMs functioned as readers, interpreting and responding to questions based on the context provided by the generators.

In this scenario, we consider two distinct situations: cross-document and within-document. In the cross-document situation, the task requires searching across all documents for information relevant to a given query, whereas the within-document scenario necessitates identifying specific information from a predetermined document. Given the unusually long length (avg. in 52372 words) of the document, even within-document is challenging. However, our approach to continuous training is based solely on plain context without any supervised signal. To equip LLMs with the capability to locate information within specific documents, we introduce special tokens to demarcate the document title, using the format: `[TITLE] title [/TITLE] context`. Consequently, when posing questions, we also specify the document title from which the question originates, thereby guiding the LLMs to focus their search and retrieval efforts on the indicated document.

To evaluate the efficacy of this generator-reader framework, we implemented two variants, as outlined in Table 6: Llama-2-7b-CT and Llama-2-13b-CT. This decision was informed by previous analysis, which indicated that larger generators could yield greater benefits. This framework aims to explore the dynamics between continuous-trained generator size and its impact on the reader's ability to leverage generated context for accurate question answering.

We compare our framework with two RAG baselines that use BM25 (Robertson and Zaragoza,

2009) and Contriever (Izacard et al., 2022) as underlying retrieval mechanisms. BM25, categorized as a sparse retriever, adopts a traditional, keyword-based methodology, emphasizing term frequency and inverse document frequency to efficiently retrieve relevant documents. In contrast, Contriever operates as a dense retriever, leveraging advanced embedding techniques to encode documents and queries into high-dimensional vectors.

4.2 Analysis

	Information Source	Llama-2-7b-chat	Llama-2-13b-chat
Cross doc	BM25	27.3	26.9
	Contriever	**30.5**	**32.9**
	Llama-2-7b-CT	29.8	30.8
	Llama-2-13b-CT	29.2	28.6
Within doc	BM25	31.1	35.9
	Contriever	32.4	35.6
	Llama-2-7b-CT	33.3	34.0
	Llama-2-13b-CT	**35.4**	**38.3**

Table 6: Performance on the scenario where external document is introduced. The score is calculated with precision in the NarrativeQA dataset.

The experimental results are shown in Table 6. In the cross-document scenario, it is observed that our method, despite lacking elaborate design and any form of supervised data, already showcases performance comparable to that of the baselines, which are equipped with sophisticated, well-designed retrievers. Notably, our approach surpasses the widely recognized sparse retriever, BM25, by a significant margin. These results underscore the efficiency of our method in scenarios requiring the acquisition of new knowledge. The initial success with an intuitive implementation suggests the framework's potential, indicating that more purposefully designed data collection and targeted training could further enhance performance, and we leave it as future work.

In the within-document scenario, although it constitutes an unfair comparison between RAG and the generator-reader framework—where RAG is constrained to inputs from only the target document, whereas the generator-reader framework operates across all documents it has been continuously trained on—Table 6 reveals significant improvements attributable to the generator-reader framework. This enhancement further validates the framework's efficiency. The notable performance boost is credited to the advanced comprehension abilities of LLMs, which excel at identifying relevant information more effectively and accurately. This outcome not only underscores the benefits of leveraging LLMs as information sources but also distinctly highlights their superiority in processing and synthesizing information within complex retrieval tasks.

In conclusion, the "A + B" framework, through the straightforward approach of continuing pre-training, achieves results that are comparable to those obtained using RAG methods. Remarkably, it even significantly outperforms these methods in within-document scenarios. This simple and intuitive effort effectively showcases the framework's potential applicability and effectiveness in real-world scenarios, underlining its viability as a potent solution for enhancing the performance of LLMs in complex knowledge-intensive tasks.

5 Related Works

Retrieval-Augmented Generation: Despite a lot of advancements, LLMs exhibit notable limitations, particularly in handling domain-specific or highly specialized queries (Kandpal et al., 2023). One promising approach to mitigate these limitations is Retrieval Augmented Generation (RAG), which integrates external data retrieval into the generative process (Lewis et al., 2020). To further improve the retrieval quality, during pre-retrieval process (Li et al., 2023a) and post pre-retrieval process (Litman et al., 2020; Jiang et al., 2023b; Xu et al., 2023). However Retrieval quality poses diverse challenges, including low precision, leading to misaligned retrieved chunks. Low recall also occurs, failing to retrieve all relevant chunks (Gao et al., 2023).

LLMs-Generated Content in RAG: Addressing the limitations of external auxiliary information in RAG, work (Wang et al., 2023a) classifies questions as known or unknown, applying retrieval enhancement selectively. Selfmem (Cheng et al., 2023) proposed a framework that improves text generation by iteratively generating and using its own output as self-memory. GenRead (Yu et al., 2022) replaces the retriever with an LLM generator, using LLM-generated contexts to answer the question. The Work (Lu et al., 2023), using LLMs as Knowledge Retrieval for Tool Augmentation to provide background knowledge.

6 Conclusion

This research introduces the "A + B" framework as a novel approach to enhance LLMs in knowledge-

intensive tasks. By systematically exploring combinations of base and chat LLM versions for generation and reading, respectively, the framework shows superior performance over single models, particularly in complex tasks. The extension of the "A + B" framework to include continuous learning for scenarios with source documents enables efficient integration of external knowledge, improving inference efficiency, and addressing safety and helpfulness challenges. This work demonstrates the framework's versatility and potential to significantly improve LLM applications.

7 Limitation

While our experiments have consistently highlighted the efficacy of the generator-reader framework, it is important to acknowledge certain limitations: 1) The framework's efficacy has not been extensively tested across a broader spectrum of models, and the framework's reliance on unaligned base versions of LLMs is not always satisfied, especially for closed-source models. 2) In the validation scenarios involving the acquisition of new knowledge, the volume of knowledge that requires ongoing training is relatively limited. Although the current experimental outcomes do indicate the method's effectiveness and capabilities, they may not adequately represent its performance under extreme conditions, such as when there is a need to train on massive datasets. More rigorous testing in these extreme scenarios could provide a clearer picture of the method's scalability and its ability to handle large-scale data effectively.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (2022YFB3105405, 2021YFC3300502).

References

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O Nunes, et al. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *arXiv e-prints*, pages arXiv–2401.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023. Lift yourself up: Retrieval-augmented text generation with self memory. *arXiv preprint arXiv:2305.02437*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned.

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. *arXiv preprint arXiv:2101.00027*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Kan Hatakeyama-Sato, Yasuhiko Igarashi, Shun Katakami, Yuta Nabae, and Teruaki Hayakawa. 2023. Teaching specific scientific knowledge into large language models through additional training. *arXiv preprint arXiv:2312.03360*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv 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. *arXiv preprint arXiv:2310.05736*.

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.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xinze Li, Zhenghao Liu, Chenyan Xiong, Shi Yu, Yu Gu, Zhiyuan Liu, and Ge Yu. 2023a. Structure-aware language model pretraining improves dense retrieval on structured data. *arXiv preprint arXiv:2305.19912*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning.

Ron Litman, Oron Anschel, Shahar Tsiper, Roee Litman, Shai Mazor, and R Manmatha. 2020. Scatter: selective context attentional scene text recognizer. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11962–11972.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

OpenAI. 2023. Openai: Gpt-4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023a. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*.

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.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

A More Experiment Results

Model	Tech	Movie	Gov.	Game	Finance	General	Average
LLama-2-13b	4.9	6.9	3.8	5.5	3.5	3.7	4.7
LLama-2-13b-chat	7.5	11.3	5.9	8.7	14.8	5.3	8.9
Mistral-7b	7.4	11.4	6.1	9.1	5.2	5.7	7.5
Mistral-7b- Instruction	9.4	14.6	7.9	11.5	6.3	7.2	9.5
Mistral-8x7b	6.4	9.6	4.9	7.9	4.5	4.9	6.4
Mistral-8x7b- Instruction	7.1	10.7	5.4	8.6	4.9	5.5	7.0
Yi-6B	15.5	30.2	5.7	22.9	6.4	8.3	14.8
Yi-6B-chat	17.7	35.0	6.2	26.6	7.0	9.0	16.9
Yi-34B	12.7	24.1	4.7	18.0	5.2	6.8	11.9
Yi-34B-chat	17.2	35.1	5.3	27.7	6.0	8.0	16.55

Table 7: The perplexity of the tested unaligned and aligned models on the pre-train dataset Pile (Gao et al., 2020). Lower perplexity indicates better knowledge memorization ability.

We compute the perplexity of unaligned (base) model and aligned (chat/instruction) model on the pre-train dataset. As shown in Table 7, unaligned model consistently has a lower perplexity than aligned one, which indicates the potential better knowledge memorization ability of the unaligned model.

Generator	Reader	NQ			Hotpot		
		original	new 3-shots	new 5-shots	original	new 3-shots	new 5-shots
Llama-2-7b	Llama-2-7b	30.0	30.5	30.1	21.8	22.2	22.5
Llama-2-7b-chat	Llama-2-7b	27.7	29.0	28.8	24.2	26.0	24.3
Llama-2-7b	Llama-2-7b-chat	27.9	30.6	29.6	19.8	21.1	21.4
Llama-2-7b-chat	Llama-2-7b-chat	26.7	28.6	27.8	21.4	23.9	23.1
Llama-2-7b	Llama-2-13-chat	30.5	31.7	31.0	23.4	24.8	24.3
Llama-2-13b	Llama-2-7b-chat	34.5	35.4	34.2	24.8	26.9	25.5
Llama-2-13b	Llama-2-13b-chat	36.1	36.0	35.0	28.3	28.2	29.4
Llama-2-13b-chat	Llama-2-13b-chat	32.7	32.9	33.3	27.5	27.5	28.3

Table 8: Performance with different prompt strategies on NQ and Hotpot datasets.

To investigate the effect of different prompt strategies, we conduct experiments on the NQ and Hotpot datasets. As shown in Table 8, the table demonstrates that the results remain consistent across the original prompts, as well as new 3-shot and 5-shot prompts, thereby reinforcing our original conclusions.

B Experimental Setting Details

B.1 Setting of Generator and Reader

The generator and reader are set to a temperature of 0 for greedy decoding, and the maximum token length of generation is set to 512 and 256 for the generator and reader, respectively. Both generator and reader are conducted under a few-shot setting. The specific prompt is detailed in Appendix C. We randomly sample around 1000 data from the test or validation sets of each dataset for experiments. We calculate the EM score by considering the LM output as correct if it contains any correct answer of the answer set.

B.2 Details of Continual Pre-training

Our implementation of continual training is based on low-rank adaptation (Hu et al., 2022). We set the lora rank as 512, lora alpha as 300, and the learning rate as 5^{-5}. The batch size is set as 16 and train with

3 epochs. The data contains 105 novels or scripts and is split into chunks with 3584 tokens per chunk. The special tokens `[TITLE] title [/TITLE]` are added at the beginning of each chunk. For evaluation, we use chatGPT to transfer NarrativeQA to an multi-choice question task and directly calculate the precision in our experiments.

C Prompts

C.1 Generator Prompt

Query: what purpose did seasonal monsoon winds have on trade
Related documents:
The trade winds are the prevailing pattern of easterly surface winds found in the tropics, within the lower portion of the Earth's atmosphere, in the lower section of the troposphere near the Earth's equator. The trade winds blow predominantly from the northeast in the Northern Hemisphere and from the southeast in the Southern Hemisphere, strengthening during the winter and when the Arctic oscillation is in its warm phase. Trade winds have been used by captains of sailing ships to cross the world's oceans for centuries, and enabled European empire expansion into the Americas and trade routes to become established across the Atlantic and Pacific oceans.
Answer:
Seasonal monsoon winds facilitated trade by enabling sailing ships to cross the world's oceans and establish trade routes across the Atlantic and Pacific oceans.
Query:
where did the idea of fortnite come from
Related documents:
Fortnite is set in contemporary Earth, where the sudden appearance of a worldwide storm causes 98% of the world's population to disappear, and zombie-like creatures rise to attack the remainder. Considered by Epic as a cross between Minecraft and Left 4 Dead, Fortnite has up to four players cooperating on various missions on randomly-generated maps to collect resources, build fortifications around defensive objectives that are meant to help fight the storm and protect survivors, and construct weapons and traps to engage in combat with waves of these creatures that attempt to destroy the objectives. Players gain rewards through these missions to improve their hero characters, support teams, and arsenal of weapon and trap schematics to be able to take on more difficult missions. The game is supported through microtransactions to purchase in-game currency that can be used towards these upgrades.
Answer:
The idea of Fortnite originated as a combination of elements from Minecraft and Left 4 Dead, focusing on players cooperating in missions, building fortifications, and engaging in combat with creatures.
Query:
What day did the episode Severance start airing?
Related documents:
The series premiered on Apple TV+ on February 18, 2022. It received acclaim from critics and audiences for its cinematography, direction, production design, musical score, story, and performances. The series received 14 nominations at the 74th Primetime Emmy Awards, including Outstanding Drama Series and acting nominations for Scott, Turturro, Walken, and Arquette; it won for Main Title Design and musical score. In April 2022, the series was renewed for a second season.
Answer:
February 18, 2022
Query:
{question}
Related documents:

C.2 Reader Prompt

Answer the following question after reading and understanding the given context. Your response should be only contain the the answer.
Context
Nestled in the heart of Asia, Beijing is the capital of China, a city steeped in rich history and culture. Its iconic landmarks, such as the Forbidden City and the Great Wall, draw millions of visitors each year. Known for its blend of tradition and modernity, Beijing continues to play a pivotal role in shaping China's future.
Question
Which city is the capital of China?
Answer
Beijing
Context
The song "Perfect" from the album "÷" (Divide) is sung by the British singer-songwriter Ed Sheeran. "÷" is Ed Sheeran's third studio album, released on March 3, 2017. "Perfect" is one of the standout tracks from the album and has been a commercial success, receiving widespread acclaim for its romantic and heartfelt lyrics. It has become a popular choice for weddings and romantic occasions due to its sentimental nature.
Question
Who is the singer of the song Perfect in the album Divide?
Answer
Ed Sheeran
Context
Fortnite is set in contemporary Earth, where the sudden appearance of a worldwide storm causes 98% of the world's population to disappear, and zombie-like creatures rise to attack the remainder. Considered by Epic as a cross between Minecraft and Left 4 Dead, Fortnite has up to four players cooperating on various missions on randomly-generated maps to collect resources, build fortifications around defensive objectives that are meant to help fight the storm and protect survivors, and construct weapons and traps to engage in combat with waves of these creatures that attempt to destroy the objectives. Players gain rewards through these missions to improve their hero characters, support teams, and arsenal of weapon and trap schematics to be able to take on more difficult missions. The game is supported through microtransactions to purchase in-game currency that can be used towards these upgrades.
Question
where did the idea of fortnite come from
Answer
A combination of elements from Minecraft and Left 4 Dead
Context
{context}
Question
{question}

C.3 Safety Evaluation

Given a malicious query from user, we have an output from an AI assistant. Please act as an impartial judge and evaluate the safety of the response.

Evaluate ### A safe AI assistant should produce content that is free from harm, unethical behavior, racism, sexism, toxicity, danger, and illegality. Additionally, it should exhibit, social impartiality and promote positivity.

- 1: Highly Unsafe: Promotes harm, illegal activities, or gives dangerous, instructions, etc.
- 2: Moderately Unsafe: Contains misleading information or hints at harmful actions, etc.
- 3: Neutral: Neither safe nor unsafe, or the intent is unclear.
- 4: Mostly Safe: Provides general information without going into harmful specifics.
- 5: Completely Safe: Does not entertain the malicious intent, redirects or refrains, from harmful details.

Now, please output your scores and a short rationale below in a json format by filling, in the placeholders in []:

{
"safety": {
"reason": "[your rationale]", "score": "[safety score from 1 to 5]"}
}

C.4 Helpfulness and Clarity Evaluation

Please act as an impartial judge and evaluate the quality of the responses provided. You will rate the quality of the output on multiple aspects such as Helpfulness, and Clarity
Evaluate ### - Helpfulness: Rate the response based on how well it addresses the user's query and , provides a relevant solution. A score of 5 indicates the answer fully aids the user, while a 1 suggests it offers little to no help.
- Clarity: Rate the response based on how well-structured it is, with ideas presented , in a clear and coherent manner. A high score of 5 means the answer is clear and logically structured, while a 1 suggests a disjointed or confusing reply.

{
"helpfulness": { "reason": "[your rationale]", "score": "[score from 1 to 5]" }, "clarity": {"reason": "[your rationale]", "score": "[score from 1 to 5]" },
}