Thread of Thought Unraveling Chaotic Contexts

Yucheng Zhou[1]*, **Xiubo Geng**[2], **Tao Shen**[3], **Chongyang Tao**[2],
Guodong Long[3], **Jian-Guang Lou**[2]†, **Jianbing Shen**[1]†
[1] SKL-IOTSC, CIS, University of Macau,
[2]Microsoft Corporation, [3]AAII, FEIT, University of Technology Sydney
yucheng.zhou@connect.um.edu.mo, {xigeng,chongyang.tao,jlou}@microsoft.com
{tao.shen, guodong.long}@uts.edu.au, jianbingshen@um.edu.mo

Abstract

Large Language Models (LLMs) have ushered in a transformative era in the field of natural language processing, excelling in tasks related to text comprehension and generation. Nevertheless, they encounter difficulties when confronted with chaotic contexts (e.g., distractors rather than long irrelevant context), leading to the inadvertent omission of certain details within the chaotic context. In response to these challenges, we introduce the "Thread of Thought" (ThoT) strategy, which draws inspiration from human cognitive processes. ThoT systematically segments and analyzes extended contexts while adeptly selecting pertinent information. This strategy serves as a versatile "plug-and-play" module, seamlessly integrating with various LLMs and prompting techniques. In the experiments, we utilize the PopQA and EntityQ datasets, as well as a Multi-Turn Conversation Response dataset (MTCR) we collected, to illustrate that ThoT significantly improves reasoning performance compared to other prompting techniques.

1 Introduction

Large Language Models (LLMs) represent a significant advancement in the field of artificial intelligence. They have achieved notable accomplishments in natural language understanding and generation (Brown et al., 2020; Wei et al., 2022). The development of LLMs has had a far-reaching impact, drawing significant attention in academia. These models demonstrate proficiency in a wide array of natural language processing tasks, including sentiment analysis (Zhang et al., 2023), machine translation (Moslem et al., 2023), and summarization (Tam et al., 2023). Moreover, they exert a profound influence across various industries and offer promising solutions for intricate issues, such as aiding in legal consultations (Yue et al., 2023) and assisting in medical diagnostics (Wang et al., 2023a).

With the growing complexity and diversity of tasks demanding extensive information processing and reasoning, particularly in the context of Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) and conversational (Xu et al., 2022) scenarios, the input text often comprises a wealth of information from various sources, including user queries, conversation history, external knowledge bases, and more. This information may be interconnected or entirely unrelated. Moreover, the significance of this information can fluctuate based on the context, with certain pieces being critical for addressing specific questions and others being extraneous. This situation can aptly be characterized as a "Chaotic Context". Similar to but distinct from "Long Context", "Chaotic Context" underscores the complexity and volume of information, going beyond the mere length of the context. Moreover, Liu et al. (2023) found that existing LLMs often encounter difficulties in effectively identifying relevant information from the context augmented through retrieval, particularly when it is located in the middle position.

Recent studies (Xu et al., 2023; Jiang et al., 2023) have proposed various solutions to enhance the performance of LLMs in long-context scenarios, upon an intuition of input capacity optimization. Xu et al. (2023) proposes a method to compare and combine retrieval-augmentation and long context extension for large language models (LLMs) on various long text tasks. However, this approach necessitates the retraining of LLMs with positional interpolation. In addition, Jiang et al. (2023) introduces LongLLMLingua, a method that streamlines the input prompt by culling irrelevant or redundant information. Nonetheless, this method mandates the fine-tuning of auxiliary models (e.g., LLaMA-7B (Touvron et al., 2023a))

*Work is done during internship at Microsoft.
†Corresponding author.

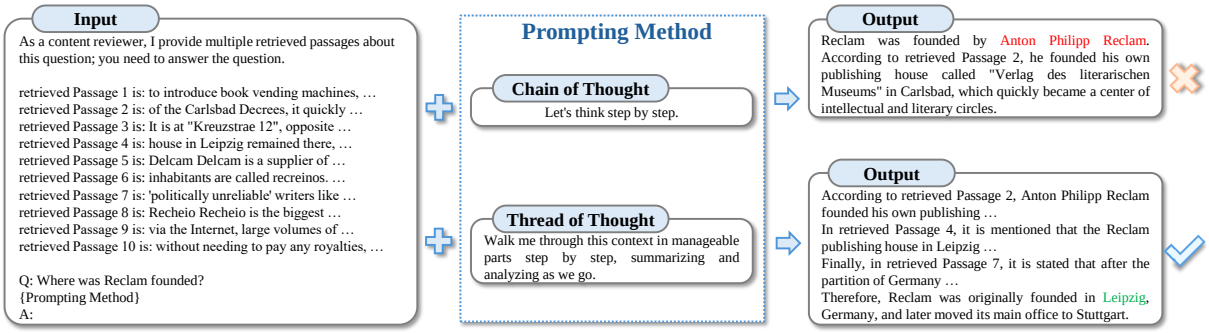


Figure 1: Thread of Thought prompting enables large language models to tackle chaotic context problems. In the output depicted, green text denotes the correct answer, while red text indicates the erroneous prediction.

for prompt compression. The utility of these auxiliary models may prove insufficient in addressing unknown or intricate content, and it imposes limitations on the length of text that can be effectively processed. Moreover, its non-end-to-end framework can lead to error propagation. In contrast, chain-of-thought (CoT) prompting (Wei et al., 2022) can enhance a model's reasoning ability without requiring any retraining or fine-tuning of LLMs. However, due to the massive amount of information contained within chaotic contexts, CoT still encounters information missing in reasoning, as shown in Figure 1.

To address these challenges, we introduce the "Thread of Thought" (ThoT) strategy. ThoT, drawing inspiration from human cognitive processes, enables Large Language Models (LLMs) to methodically segment and analyze extended contexts. This segmentation enhances the extraction of pertinent content for responding to queries. ThoT represents the unbroken continuity of ideas that individuals maintain while sifting through vast information, allowing for the selective extraction of relevant details and the dismissal of extraneous ones. This balance of attention across a document's sections is crucial for accurately interpreting and responding to the information presented. Moreover, the stepwise analysis and summarization of segmented information improve comprehension over multiple paragraphs and protect LLMs against misleading yet seemingly relevant data.

In comparison to existing methods that require complex multi-stage prompting (Zhou et al., 2023) or multi-path sampling (Wang et al., 2023b), ThoT is a simpler, more universal, and efficient solution. It integrates seamlessly as a "plug-and-play" module with various pre-trained language models and prompting strategies, avoiding complex procedures. ThoT not only improves LLMs' performance in chaotic contexts but also enhances their reasoning abilities.

To evaluate ThoT's effectiveness in handling chaotic contextual information, we used long-tail question answering datasets, specifically PopQA (Mallen et al., 2023) and EntityQ (Sciavolino et al., 2021). These datasets feature knowledge often unfamiliar to large models, thereby reducing the impact of their inherent knowledge retention on our results. Additionally, we construct a Multi-Turn Conversation Response (MTCR) dataset based on everyday conversations to further assess our method. Comparative analyses with other prompting techniques show that ThoT markedly improves reasoning performance, evidencing its effectiveness. We also explored various prompts to determine optimal prompting strategies.

2 Related Work

2.1 Long Context Large Language Models

Recent advancements in Large Language Models (LLMs) have made significant strides in managing extended contexts, moving beyond the limitations of traditional pre-defined context windows. Ratner et al. (2023) introduce the Parallel Context Windows (PCW) method, which segments extensive contexts into multiple windows, employing independent attention mechanisms. Building on this concept, Chen et al. (2023) facilitate substantially longer context windows with minimal fine-tuning by aligning position indices with the maximum position index from the pre-training phase. Moreover, a different approach, LongNet, utilizes dilated attention, allowing the attention field to expand exponentially with distance (Ding et al., 2023). In addition, Xiao et al. (2023) underscore the phenomenon of attention convergence, where maintaining the

Key-Value (KV) states of initial tokens significantly enhances window attention performance. Lastly, Press et al. (2022) introduce Attention with Linear Biases (ALiBi), a method that biases the query-key attention scores based on distance, achieving comparable perplexity to models trained on longer sequences. However, these methods predominantly concentrate on long contexts. In contrast, chaotic contexts are characterized by their overloaded information, often cluttered with numerous similar and unrelated elements.

2.2 Reasoning with Large Language Models

Advancements in large language models (LLMs) have significantly impacted AI, notably in complex reasoning tasks. The enhancement of LLMs' reasoning capabilities is exemplified in (Wei et al., 2022), where Chain-of-Thought (CoT) prompting is introduced. This method improves arithmetic, common sense, and symbolic reasoning by generating intermediate steps. Building on this, the Graph of Thoughts (GoT) framework conceptualizes LLM outputs as graphs, leading to notable improvements in task performance and efficiency (Besta et al., 2023). Extending the CoT concept, Yao et al. (2023a) propose the Tree of Thoughts (ToT) framework, which has shown remarkable success in complex problem-solving tasks like the 24-point game. In addition, Zhou et al. (2023) introduce the least-to-most prompting strategy, breaking down complex problems into simpler subproblems and showing effectiveness in tasks requiring advanced symbolic manipulation. Lastly, Yao et al. (2023b) explore non-linear thought processes through GoT reasoning, outperforming the linear CoT approach in both mathematical and financial problem datasets. However, these methods are effective but overlook chaotic context scenarios.

2.3 Knowledge Following in Long Context

LLMs can process extensive input contexts, but their performance significantly deteriorates when extracting relevant information buried in these contexts, challenging their efficiency in managing long contexts (Liu et al., 2023). To address deploying LLMs in streaming applications, Xiao et al. (2023) introduce the StreamingLLM framework, enabling LLMs with limited attention windows to handle indefinitely long sequences without additional fine-tuning. Some study finds that retrieval augmentation enables a 4K context window LLM to equal the performance of a 16K context window LLM fine-tuned with positional interpolation in long-context tasks, underscoring the potential of retrieval methods in augmenting LLM capabilities (Xu et al., 2023). Moreover, LongLLMLingua introduces prompt compression to improve LLMs' key information perception, significantly boosting performance (Jiang et al., 2023).

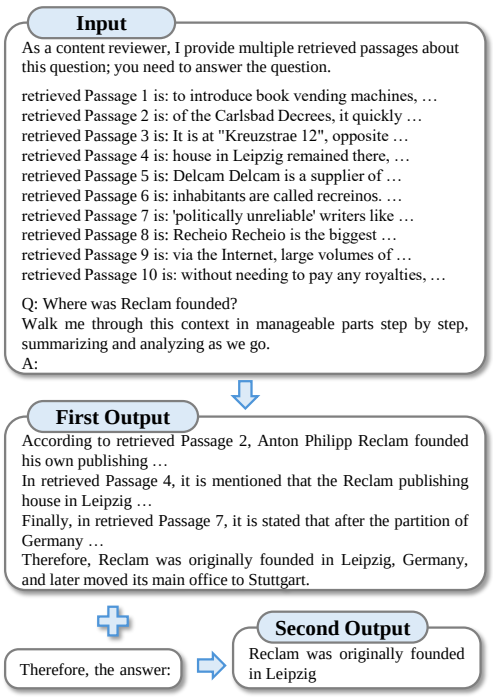


Figure 2: Thread of Thought for zero-shot reasoning.

3 Methodology

We present an innovative method for template-based prompting that is specifically designed to enhance Thread of Thought (ThoT) reasoning. This novel strategy stands distinct from the traditional chain of thought prompting (Wei et al., 2022), adept at navigating through disordered contexts in which the information may be either interwoven or disparate. ThoT prompting can be seamlessly integrated with a variety of existing language models and prompting techniques, offering a modular "plug-and-play" improvement that eliminates the need for elaborate prompting strategies or sampling methods. Our approach's underlying principle is both simple and efficient, as exemplified in Figure 2: inserting "Walk me through this context in manageable parts step by step, summarizing and analyzing as we go" into the prompt facilitates ThoT reasoning.

As illustrated in Figure 2, in contrast to Chain

of Thought (CoT) prompting, which struggles with complex and chaotic contexts, ThoT prompting adeptly maintains the logical progression of reasoning without being overwhelmed. While prompt compressors and similar strategies have sought to address these complexities, they often underperform with unfamiliar or particularly complex material and typically necessitate significant modifications to the Large Language Models (LLMs), such as retraining or fine-tuning with additional datasets (Xu et al., 2023; Jiang et al., 2023). ThoT, however, not only effectively manages chaotic contexts but also simplifies the prompting process, requiring just two prompting efforts compared to CoT.

3.1 First Step: Initiating the Reasoning

The initial prompt is designed to guide the LLM through an analytical dissection of the context, using the directive "Walk me through this context in manageable parts step by step, summarizing and analyzing as we go". Specifically, we employ a template that incorporates the chaotic context $\mathbb{X}$ and query $\mathbb{Q}$ into the prompt $\mathbb{P}$ as "[$\mathbb{X}$] Q: [$\mathbb{Q}$] [$\mathbb{T}$] A:", where [$\mathbb{T}$] denotes the trigger sentence $\boldsymbol{t}$ that initiates the reasoning process. For instance, utilizing "Walk me through this context in manageable parts step by step, summarizing and analyzing as we go" as the trigger, the prompt $\mathbb{P}$ becomes "[$\mathbb{X}$] Q: [$\mathbb{Q}$] Walk me through this context in manageable parts step by step, summarizing and analyzing as we go. A:". This prompted text $\mathbb{P}$ is then inputted into an LLM, which generates the subsequent sentences $\mathbb{Z}$. This procedure is modeled after the cognitive strategies humans employ when confronted with complex information, breaking it down into digestible segments, distilling key points, and navigating through the material with sustained focus. This incremental method fosters a more structured and coherent line of reasoning, proving particularly advantageous in chaotic contexts.

3.2 Second Step: Refining the Conclusion

The second prompt builds upon the structured reasoning established earlier, employing another prompt to distill the analysis into a definitive answer. By leveraging the organized thought sequence initiated by the first prompt, this step aims to succinctly capture the essence of the conclusion. Specifically, we use a simple template to combine the initial prompted text $\mathbb{P}$, the response $\mathbb{Z}$, and the conclusion marker [$\mathbb{A}$], as in "[$\mathbb{P}$] [$\mathbb{Z}$] [$\mathbb{A}$]", where [$\mathbb{A}$] signifies the trigger sentence designed to extract the answer, such as "Therefore, the answer:". This extraction prompt perpetuates the thought process, prompting the model to sift through the analysis and isolate the principal conclusion as the final answer. The prompt's design is a deliberate tactic to sharpen the model's focus, fostering precision and explicitness in the response.

This two-tiered prompting system effectively addresses the limitations of prior methods while obviating the need for intensive model retraining or complex modifications. Our methodology not only enhances the model's capacity to navigate chaotic contexts but also more closely aligns its reasoning processes with human cognitive patterns.

Generate Responds Based on Conversation and Persona

As a writer, your task is to continue the following dialogue. Start by inferring the possible current situation of Speaker 2 based on their persona. Then, have Speaker 1 ask a question that indirectly alludes to Speaker 2's situation without directly mentioning the words from their persona. Finally, Speaker 2 should respond, incorporating his persona into the answer.

These are the conversations of the two speakers:
{Conversation}

These are the personas of the speaker 2:
{Speaker2's personas}

Please continue the above conversation, with one persona, one response from Speaker 1 and one response from Speaker 2. Think step-by-step, then explain.

Extract Responds From Generated Conversation

The final output includes the persona and dialogue between speaker 1 and speaker 2 in the following format:
Persona: {persona}
S1: {response1}
S2: {response2}

Figure 3: Prompt for MTCR Dataset Construction.

4 Experiments

4.1 Experimental Settings

Dataset. We evaluated our method across two chaotic context scenarios: retrieval-augmented generation and multi-turn conversation response. Our assessment utilized three datasets: the PopQA dataset (Mallen et al., 2023), the EntityQ dataset (Sciavolino et al., 2021), and our own Multi-Turn Conversation Response (MTCR) dataset. Specifically, the PopQA and EntityQ datasets, designed to contain long-tail knowledge, were chosen to minimize interference from the extensive internal knowledge of large models, thereby facilitating a more effective comparison of different methodologies. Distinct from the original PopQA and EntityQ datasets, we randomly selected a test set of 1,000 samples for our analysis. For the evaluation of the PopQA and EntityQ datasets, we adhered to the original datasets' metric, namely the exact match (EM). Furthermore, the MTCR dataset, used to assess multi-turn conversation response, was de-

Generate Responds Based on Conversation and Persona

Assessment of the Quality of Generated Speaker2's Response

Conversation Content:
{conversation}

Generated Respond:
{generated speaker2's response}

Persona:
{speaker2's persona}

Comprehensive Evaluation Guide:
Score each of the following three criteria separately.

Relevance:
- 1 point: Not relevant; the response does not relate to Speaker1's dialogue.
- 2 points: Slightly relevant; the response touches on the subject but misses key points or deviates significantly.
- 3 points: Somewhat relevant; the response is related to Speaker1's dialogue but may miss some nuances or details.
- 4 points: Relevant; the response is on topic and addresses most points made by Speaker1.
- 5 points: Highly relevant; the response is fully on topic, directly addresses all elements of Speaker1's dialogue.

Accuracy:
- 1 point: Inaccurate; the response contains significant errors or shows misunderstanding of the topic.
- 2 points: Somewhat inaccurate; the response contains multiple errors, though it grasps the basic idea.
- 3 points: Moderately accurate; the response has minor errors but generally understands the topic.
- 4 points: Mostly accurate; the response contains minimal, inconsequential errors.
- 5 points: Fully accurate; the response is free from errors and fully understands the topic.

Persona Representation:
- 1 point: No representation; Speaker2's persona is not reflected in the response.
- 2 points: Weak representation; Speaker2's persona is hinted at but largely absent or incorrect.
- 3 points: Adequate representation; Speaker2's persona is present but some traits may be missing or not fully captured.
- 4 points: Strong representation; Speaker2's persona is clear and most traits are well represented.
- 5 points: Full representation; Speaker2's persona is fully and accurately portrayed throughout the response.

Example Output Form:

Score:
Relevance Score: {score}
Accuracy Score: {score}
Persona Representation Score: {score}

Scoring Rationale:
Relevance Score: {scoring rationale}
Accuracy Score: {scoring rationale}
Persona Representation Score: {scoring rationale}

Figure 4: Prompt Evaluation Metric for MTCR Dataset.

veloped based on the Multi-Session Chat (MSC) dataset (Xu et al., 2022). The dataset construction involved sequentially using two prompts, as shown in Figure 3. The input of prompts is the MSC dataset's conversation and Speaker2's persona to generate a response for Speaker1. During the inference phase, the model was required to consider the multi-turn conversation contextual details mentioned previously to generate a response for speaker2, coping with the response created for speaker1. Following this, a manual screening process was conducted to eliminate samples that did not meet certain criteria, such as persona content leakage and irrelevance to the context or persona, culminating in a refined selection of 304 samples. For the MTCR dataset's evaluation, we merge the persona as a known condition along with the model-generated response for Speaker2 in the prompt, as depicted in Figure 4, and then pass them into GPT-4 (OpenAI, 2023), obtaining scoring.

Prompt. In the experimental comparison, we consider four distinct prompts for retrieval-augmented generation. (1) "Vanilla" entails using the instruction and question as the prompt without providing any retrieval results, i.e., "{instruction} {question}.". (2) "Retrieval" includes retrieval results within the prompt, formatted as "{instruction} {retrieval results} {question}.". (3) "CoT" (Chain of Thought) incorporates the retrieval results and appends the phrase "Let's think step by step" to the instruction and question, resulting in "{instruction} {retrieval results} {question} Let's think step by step.". (4)"ThoT" (Thought-by-Thought) also integrates retrieval results and follows a more detailed prompt structure: "{instruction} {retrieval results} {question} Walk me through this context in manageable parts step by step, summarizing and analyzing as we go.". For the MTCR dataset, we employ only the "Vanilla", "CoT", and "ThoT" prompts. Their formats are, respectively: "{instruction} {conversation}", "{instruction} Let's think step by step. {conversation}", and "{instruction} Walk me through this context in manageable parts step by step, summarizing and analyzing as we go. {conversation}".

Method	GPT-3.5-turbo	LLaMA 2 Chat (70B)
Vanilla	0.398	0.330
Retrieval	0.475	0.510
CoT	0.482	0.525
ThoT	**0.574**	**0.561**

Table 1: Performance Comparison on PopQA.

Method	GPT-3.5-turbo	LLaMA 2 Chat (70B)
Vanilla	0.497	0.430
Retrieval	0.512	0.522
CoT	0.517	0.547
ThoT	**0.565**	**0.559**

Table 2: Performance Comparison on EntityQ.

Language models. We evaluated four large-scale language models: GPT-3.5-turbo (Schulman et al., 2022), GPT-4 (OpenAI, 2023), LLaMA 2 Chat (Touvron et al., 2023b), and Vicuna (Chiang et al., 2023). Due to the GPT-3.5-turbo and GPT-4 are not open-source, the details of their model parameters remain undisclosed. For the LLaMA 2 Chat model, we utilized variants with 7B, 13B, and 70B parameters in our experiments. Similarly, versions with 7B, 13B, and 33B parameters of the Vicuna model were employed. Sampling from these models was conducted using a greedy decoding strategy.

Method	GPT-3.5-turbo				LLaMA 2 Chat (70B)			
	Relevance	Accuracy	Persona	Average	Relevance	Accuracy	Persona	Average
Vanilla	3.211	3.135	3.345	3.230	2.819	2.901	2.914	2.878
CoT	3.352	3.220	3.349	3.307	2.783	2.806	2.882	2.823
ThoT	**3.849**	**3.921**	**3.645**	**3.805**	**3.158**	**3.295**	**3.268**	**3.240**

Table 3: Performance Comparison on MTCR dataset.

Method	PopQA			EntityQ		
	GPT-4	GPT-3.5-turbo	LLaMA 2 Chat (70B)	GPT-4	GPT-3.5-turbo	LLaMA 2 Chat (70B)
Vanilla	0.430	0.391	0.314	0.405	0.405	0.369
Retrieval	0.360	0.477	0.430	0.571	0.560	0.643
CoT	0.442	0.465	0.558	0.560	0.583	0.667
ThoT	**0.651**	**0.674**	**0.663**	**0.643**	**0.667**	**0.702**

Table 4: Study of "Lost in Middle" in PopQA and EntityQ.

4.2 Results

Tables 1 and Tables 2 show the performance of retrieval-augmented generation. In PopQA and EntityQ datasets, we notice a consistent pattern where the Thought-by-Thought (ThoT) prompt configuration outperforms the other methods. The introduction of CoT also demonstrates a positive effect, indicating that prompting models to follow a methodical problem-solving approach can improve performance metrics. It is particularly noteworthy that ThoT exhibits a marked improvement in results over the CoT configuration, highlighting the efficacy of stepwise contextual processing in enhancing the quality of generated responses. In Tables 3, a similar trend emerges. ThoT retains its lead, suggesting that its detailed prompt structure, which encourages summarizing and analyzing information in a structured manner, is particularly effective in complex conversational contexts. It underscores the importance of a methodical breakdown of context in generating relevant, accurate, and persona-consistent responses. The structured approach of ThoT prompts, which guide the model through a detailed, step-by-step analysis, consistently yields the best performance across chaotic contexts.

4.3 Lost in Middle

As shown in Table 4, we delves into the phenomena termed "Lost in Middle" (Liu et al., 2023), where the focus is to examine the performance of various models on two different question-answering datasets, PopQA and EntityQ. The presented results draw a comparison between four methodologies: Vanilla, Retrieval, Chain of Thought (CoT), and Theory of Mind (ThoT), as applied to three advanced language models: GPT-4, GPT-3.5-turbo, and LLaMA 2 Chat (70B).

Performance on PopQA : The results indicate that ThoT significantly outperforms the other methods across all three models. With GPT-4 leading at a score of 0.651, closely followed by GPT-3.5-turbo and LLaMA 2 Chat (70B) at 0.674 and 0.663, respectively. This suggests that ThoT's advanced technique, potentially incorporating more nuanced understandings of context and reasoning, has a definitive edge in handling the complexities of PopQA. The Vanilla approach yields moderate performance with GPT-4, which surpasses the scores of the other two models, hinting at the superior reasoning capabilities of the latest model iteration.

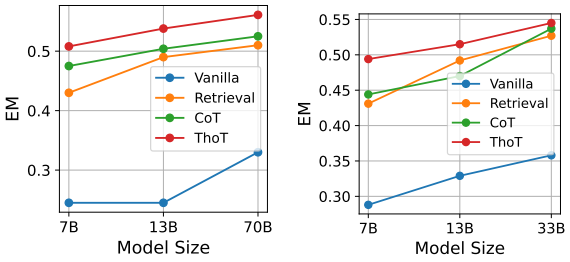
Performance on EntityQ : Similar to PopQA, the ThoT methodology again tops the charts, indicating its robustness across different datasets. GPT-4's performance, while still the highest in the Vanilla method, sees a significant jump to 0.643 when applying ThoT, suggesting a better synergy between GPT-4's capabilities and ThoT's advanced reasoning framework. Notably, the Retrieval method showcases a stark improvement over Vanilla for all models, with LLaMA 2 Chat (70B) achieving the highest score of 0.643.

4.4 Impact of Model Scale

As shown in Figure 5, results demonstrate a clear correlation between the scale of the model and its performance across different prompting strategies. As we scale up from 7 billion parameters to 70 billion parameters in the LLama2, there is a notice-

No.	Template	EM
1	Let's read through the document section by section, analyzing each part carefully as we go.	0.43
2	Take me through this long document step-by-step, making sure not to miss any important details.	0.47
3	Divide the document into manageable parts and guide me through each one, providing insights as we move along.	0.51
4	Analyze this extensive document in sections, summarizing each one and noting any key points.	0.47
5	Let's go through this document piece by piece, paying close attention to each section.	0.50
6	Examine the document in chunks, evaluating each part critically before moving to the next.	0.49
7	Walk me through this lengthy document segment by segment, focusing on each part's significance.	0.52
8	Let's dissect this document bit by bit, making sure to understand the nuances of each section.	0.45
9	Systematically work through this document, summarizing and analyzing each portion as we go.	0.45
10	Navigate through this long document by breaking it into smaller parts and summarizing each, so we don't miss anything.	0.48
11	Let's explore the context step-by-step, carefully examining each segment.	0.44
12	Take me through the context bit by bit, making sure we capture all important aspects.	0.49
13	Let's navigate through the context section by section, identifying key elements in each part.	0.47
14	Systematically go through the context, focusing on each part individually.	0.46
15	Let's dissect the context into smaller pieces, reviewing each one for its importance and relevance.	0.47
16	Analyze the context by breaking it down into sections, summarizing each as we move forward.	0.49
17	Guide me through the context part by part, providing insights along the way.	0.52
18	Examine each segment of the context meticulously, and let's discuss the findings.	0.44
19	Approach the context incrementally, taking the time to understand each portion fully.	0.42
20	Carefully analyze the context piece by piece, highlighting relevant points for each question.	0.47
21	In a step-by-step manner, go through the context, surfacing important information that could be useful.	0.53
22	Methodically examine the context, focusing on key segments that may answer the query.	0.45
23	Progressively sift through the context, ensuring we capture all pertinent details.	0.46
24	Navigate through the context incrementally, identifying and summarizing relevant portions.	0.48
25	Let's scrutinize the context in chunks, keeping an eye out for information that answers our queries.	0.42
26	Take a modular approach to the context, summarizing each part before drawing any conclusions.	0.47
27	Read the context in sections, concentrating on gathering insights that answer the question at hand.	0.48
28	Proceed through the context systematically, zeroing in on areas that could provide the answers we're seeking.	0.49
29	Let's take a segmented approach to the context, carefully evaluating each part for its relevance to the questions posed.	0.39
30	Walk me through this context in manageable parts step by step, summarizing and analyzing as we go.	**0.55**

Table 5: Prompt Selection Analysis.



(a) LLaMA 2 (b) Vicuna

Figure 5: PopQA performance on different scales of LLaMA 2 and Vicuna.

able increase in the EM scores across all prompt configurations. The same trend is observed with the Vicuna model. This increase corroborates the hypothesis that larger models have a greater capacity for understanding and generating accurate responses. The Thought-by-Thought (ThoT) configuration maintains a lead in performance at all model sizes, which demonstrates ThoT prompts appear to leverage the model's capabilities more efficiently, regardless of model size. In addition, the experimental data suggests that model scale positively impacts the performance of retrieval-augmented generation methods, with prompts such as CoT and ThoT showing the most significant benefits. This trend underscores the importance of prompt design in harnessing the full potential of large language models.

4.5 Prompt Selection

The prompt analysis, as depicted in Table 5, reveals strategic variations in phrasing that have a direct impact on the performance of the language models, as indicated by the Exact Match (EM) scores. The prompts are designed to guide the model through the process of analyzing a document or context systematically and step by step. A common theme among the more successful prompts, particularly those with EM scores above 0.50, is the explicit instruction to the model to not only analyze each part of the document but also to summarize and note key insights as it progresses. For instance, prompts that instructed the model to sum-

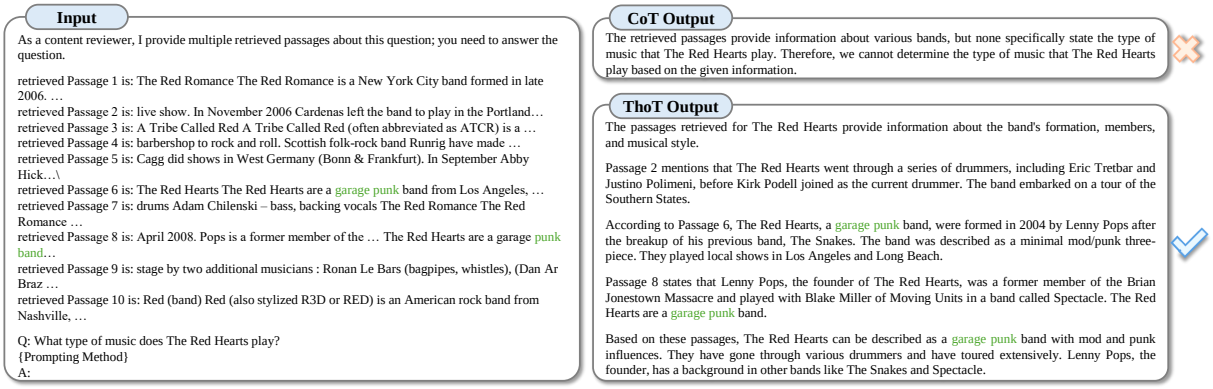


Figure 6: Case Study.

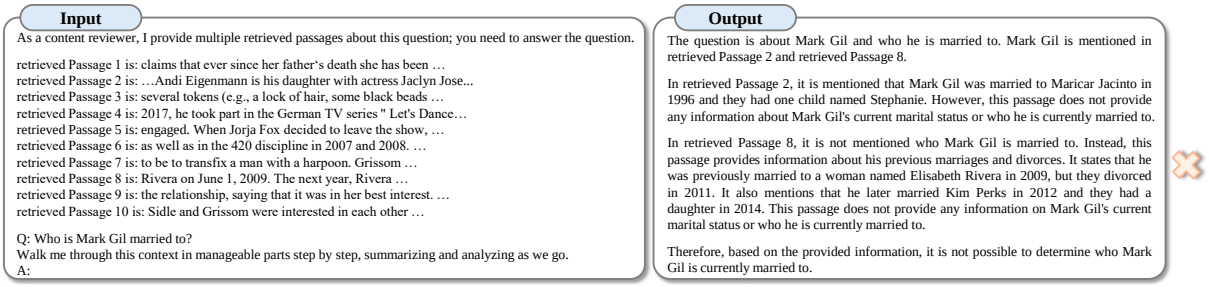


Figure 7: Error Analysis.

marize each section and not miss important details, such as prompt 2 and prompt 4, resulted in higher EM scores. Prompts that encouraged a more granular approach, directing the model to focus on individual parts and their significance or relevance, also performed well. This is evidenced by prompt 14, which achieved a relatively high EM score. The more detailed the instruction for the model to dissect and analyze the context, the better the model performed. Conversely, prompts that were less directive or less structured, such as prompt 29, tended to result in lower EM scores. This suggests that models benefit from clear, specific, and action-oriented instructions that leave little room for ambiguity in the analytical process. The highest-scoring prompt, number 30, combines several elements of successful prompts. It asks the model to manage the complexity by breaking it down into parts, which implies a thorough analysis, and also to summarize and analyze, indicating an active engagement with the material that goes beyond mere reading or passive understanding. In summary, the results suggest that prompts that are structured to enforce a detailed analytical process, encouraging step-by-step dissection, summarization, and critical evaluation, lead to better model performance.

4.6 Case Study

The case study presented in Figure 6 shows a comparative analysis between the CoT and ThoT in PopQA. CoT only stated that the passages contained information about various bands without specifying the genre of "The Red Hearts". This illustrates a potential limitation of the CoT approach: it might not effectively synthesize information from multiple sources when the answer is not explicitly stated but rather needs to be inferred from the given data. On the contrary, the ThoT method successfully identified that "The Red Hearts play garage punk music". This outcome showcases the strength of the ThoT approach. ThoT is adept at synthesizing and correlating information across multiple pieces of text. It pieced together relevant details from passages 6 and 8, noting that "The Red Hearts" were described as "a garage punk band".

4.7 Error Analysis

From Figure 7, the ThoT method can not conclude the answer for this case. The passage stating, "Andi Eigenmann is his daughter with actress Jaclyn Jose" holds the key to the correct inference that Mark Gil was married to Jaclyn Jose. The ThoT method's failure to make this inference suggests that while the model is adept at extracting explicit informa-

tion, it struggles with implicit reasoning that requires understanding nuanced relationships. The oversight may be attributed to the model's inferential reasoning capabilities, specifically regarding relationship inference—a known shortcoming in large models as also identified in prior research (Berglund et al., 2023). The case study highlights the need for models to not only parse and summarize information but also engage in a level of deductive reasoning that resembles human cognition. Therefore, enhancing the model's ability to infer and reason about entity relationships is very important.

5 Conclusion

This paper presented the "Thread of Thought" (ThoT) strategy, a novel approach designed to enhance the performance of Large Language Models (LLMs) in processing chaotic contextual information. ThoT, inspired by human cognitive processes, significantly improves the ability of LLMs to segment and analyze extended contexts. We compared ThoT with existing methods, which often require complex retraining, fine-tuning, or are limited in their ability to handle large volumes of intricate information. ThoT, in contrast, offers a more straightforward and efficient solution. It acts as a "plug-and-play" module, seamlessly integrating with various pre-trained language models and prompting strategies without necessitating complex procedures. The effectiveness of ThoT was rigorously tested using long-tail question answering datasets, such as PopQA and EntityQ, and a Multi-Turn Conversation Response dataset based on everyday conversations. The results from these evaluations were clear: ThoT not only excelled in handling chaotic contexts but also enhanced the reasoning capabilities of LLMs.

References

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a". *CoRR*, abs/2309.12288.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. *CoRR*, abs/2308.09687.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *CoRR*, abs/2306.15595.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1, 000, 000, 000 tokens. *CoRR*, abs/2307.02486.

Huiqiang Jiang, Qianhui Wu, , Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *ArXiv preprint*, abs/2310.06839.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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. *CoRR*, abs/2307.03172.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation

with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 227–237. European Association for Machine Translation.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. Parallel context windows for large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6383–6402. Association for Computational Linguistics.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6138–6148. Association for Computational Linguistics.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5220–5255. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023a. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *CoRR*, abs/2302.07257.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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 *NeurIPS*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *CoRR*, abs/2309.17453.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5180–5197. Association for Computational Linguistics.

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. *CoRR*, abs/2310.03025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601.

Yao Yao, Zuchao Li, and Hai Zhao. 2023b. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *CoRR*, abs/2305.16582.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao,

Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *CoRR*, abs/2309.11325.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *CoRR*, abs/2305.15005.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.