Chain of Evidences and Evidence to Generate: Prompting for Context Grounded and Retrieval Augmented Reasoning

Anonymous ACL submission

Abstract

While chain-of-thoughts (CoT) prompting has revolutionized how LLMs perform reasoning tasks, its current methods and variations (e.g, Self-consistency, ReACT, Reflexion, Tree-of-Thoughts (ToT), Cumulative Reasoning (CR) etc.,) suffer from limitations like limited context grounding, hallucination/inconsistent output generation, and iterative sluggishness. To overcome these challenges, we introduce a novel mono/dual-step prompting framework built upon two unique strategies **Chain of Evidences (CoE)** and **Evidence to Generate (E2G)**. Instead of unverified reasoning claims, our innovative approaches leverage the power of "evidence for decision making" by first focusing exclusively on the thought sequences explicitly mentioned in the context which then serve as extracted evidence, guiding the LLM's output generation process with greater precision and efficiency. This simple yet potent approach unlocks the full potential of chain-of-thoughts prompting, facilitating faster, more reliable, and contextually aware reasoning in LLMs. Our framework consistently achieves remarkable results across various knowledge-intensive reasoning and generation tasks, surpassing baseline approaches with state-of-the-art LLMs. For instance, (i) on the LogiQA benchmark using GPT-4, CoE achieves a new state-of-the-art accuracy of 53.8%, surpassing CoT by 18%, ToT by 11%, and CR by 9%; (ii) CoE with PaLM-2 outperforms the variable-shot performance of Gemini Ultra by 0.9 F1 points, achieving an F1 score of 83.3 on DROP.

1 Introduction

Retrieval-augmented or context-based generation serves as a mean for leveraging relevant information, empowering large language models (LLMs) to reduce the factual errors in their generation (Asai et al., 2023a,b). However, despite the expansion in model and data size, LLMs struggle in contextual reasoning. This challenge is further amplified when dealing with retrieved information that are often long and imperfect text with distractive contents.

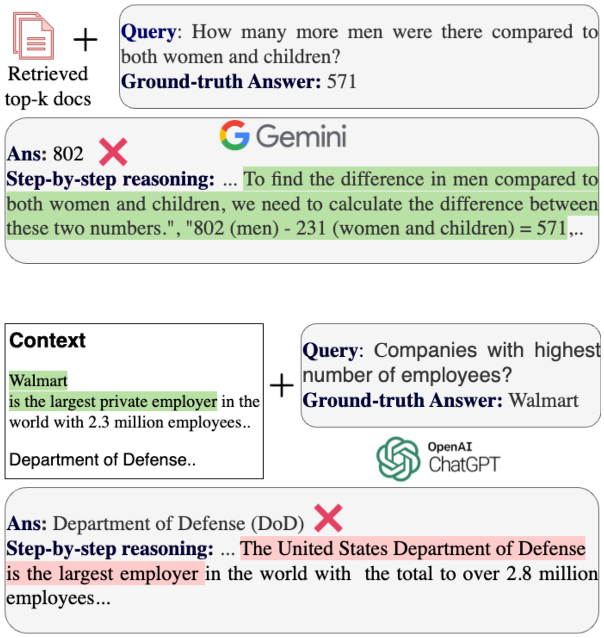

Figure 1: CoT & variants falter in context-aware reasoning. Top: Overwhelming long-text complexity leads models' failure even when it generates partially/fully correct reasoning (in green). Bottom: Ungrounded internal reasoning fails to grasp context, confusing "DoD" (ungroundeded private org in red) vs Walmart (in green).

To bolster LLM's reasoning capabilities, the Chain-of-Thought (CoT) prompting paradigm has emerged as a potent tool (Wei et al., 2022). By simulating step-by-step thinking, CoT aids in breaking down complex problems into manageable chunks. Subsequent methods, including Self-consistency (SC; (Wang et al., 2022)), ReACT (Yao et al., 2022), Reflexion (Shinn et al., 2023), Tree of Thoughts (ToT; (Yao et al., 2023)), and Cumulative Reasoning (CR; (Zhang et al., 2023b)), generalize CoT with various multi-objective, ensemble-based, or tool-augmented, and trial & error approaches but do not address the complexities of context-grounded or retrieval augmented generations (RAG). We highlight two of their pivotal bottlenecks: (i) CoT focuses solely on expanding steps without verifying hypotheses; (ii) excessively long retrieved text can lead to incorrect conclusions even with valid CoT reasonings. We depict an example in Figure 1. Additionally, challenges in finding suitable prompts or in-context exemplars for each objective, the requirement of external evaluation tools, and dependence on iterative prompting constrain their real-time document reasoning applicability. Therefore, unlocking CoT's true potential for RAG and context-aware reasoning remains unanswered. This paper delves into these challenges, exploring the limitations of existing approaches and proposing a novel prompting framework.

Ours framework consists of two unique and real-time prompting strategies particularly tailored for context-aware reasoning. First, single-step **Chain-of-Evidences (COE):** to address the problem of ungrounded reasoning hypotheses, our designed prompt asks for specific thought sequences that are explicitly mentioned in the context. We call these series of intermediate reasoning steps w/ directly extracted rationales from the given context as *evidence* (as in human decision making). Our key distinction from existing CoT approaches is that instead of mere "thinking step-by-step" (Kojima et al., 2022) our prompt instruction asks for "step-by-step reasoning w/ evidence & explanation".

Second, dual-step **Evidence to Generate (E2G):** to facilitate LLMs' answering the query properly even w/ retrieval augmented long-text contexts, we split the task into steps. In the first step (E), we adopt prompts similar to COE and generate both the $Answer$ & $Evidence$. Then in next step (G), we pass only the $Evidence$ as context for a second round of COE to LLM. G Step *Answer* is predicted as the final answer. In contrast to complex long original context in E step, the $Evidence$ is a concise short text that directly answer the input query, G step is very fast, and simpler for the model to generate answer. Additionally, in contrast to existing multi-step multi-objective CoT methods (e.g., Creswell et al. (2022); Li et al. (2023)) that employ different intermediate prompts (e.g., rationale selection & inference/premise derivation) method w/ k-shot annotated examples, using the same prompt twice in above manner, we eradicate the hurdle of choosing multiple prompts or in-context exemplars.

In experiments with multiple LLMs, we show that our prompts consistently outperform existing approaches in a diverse set of eight context-driven tasks, including natural QA, complex multi-hop, long-form QA, fact checking, dialog generation, and reading comprehension tasks. Since, even with such techniques, it is non-trivial to comprehend why and how this works and how to setup the prompt to function correctly, cost-effectively, and robustly. To this end, we perform case studies, analyze different alternatives and reveal the strengths and weaknesses of our approach. We will release the collection of our prompts and outputs on these benchmarks as a new instruction tuning dataset for future research.

2 Related Works and Preliminaries

2.1 Prompting LLMs

Various prompting paradigms have been studied in literature toward enhancing reasoning in LLMs. In Section 1, we provide a (non-exhaustive) list of CoT approaches. Among others, search-based (Pryzant et al., 2023; Lu et al., 2021), Program-aided LLM generation (Liu et al., 2023a; Gao et al., 2023; Jung et al., 2022; Zhu et al., 2022), self generation of prompts (He et al., 2023; Yasunaga et al., 2023; Sun et al., 2022; Kim et al., 2022; Li et al., 2022), self evaluation based approaches (Madaan et al., 2023; Xie et al., 2023; Kim et al., 2023; Paul et al., 2023) have been studied. Other works have also been extended w/ more complex multi-step reasoning procedure (e.g., using a different fine-tuned model (Zelikman et al., 2022; Nye et al., 2021; Lester et al., 2021)) or for domain specific applications (Parvez et al., 2023, 2021; Ouyang et al., 2022; Sanh et al., 2021; Wei et al., 2021).

2.2 Chain-of-Thoughts (CoT) Prompting

Chain-of-thoughts (CoT; (Wei et al., 2022)) is a prompting framework that guides LLMs to produce intermediate reasoning steps towards the final answer, enhancing its reasoning. Original version of CoT employs a few-shot version by providing multiple exemplars of the reasoning process (question–reasoning–answer), leveraging LLMs' in-context learning abilities. However, due to the requirement of labeled exemplars, it quickly evolved with a 0-shot instance (Kojima et al., 2022). 0-shot CoT prompts LLMs with a general instruction like "think step by step" to produce intermediate reasoning steps (See Figure 2).

3 Our Prompting Framework

In this section, we develop our prompting framework for context-grounding and retrieval aug-

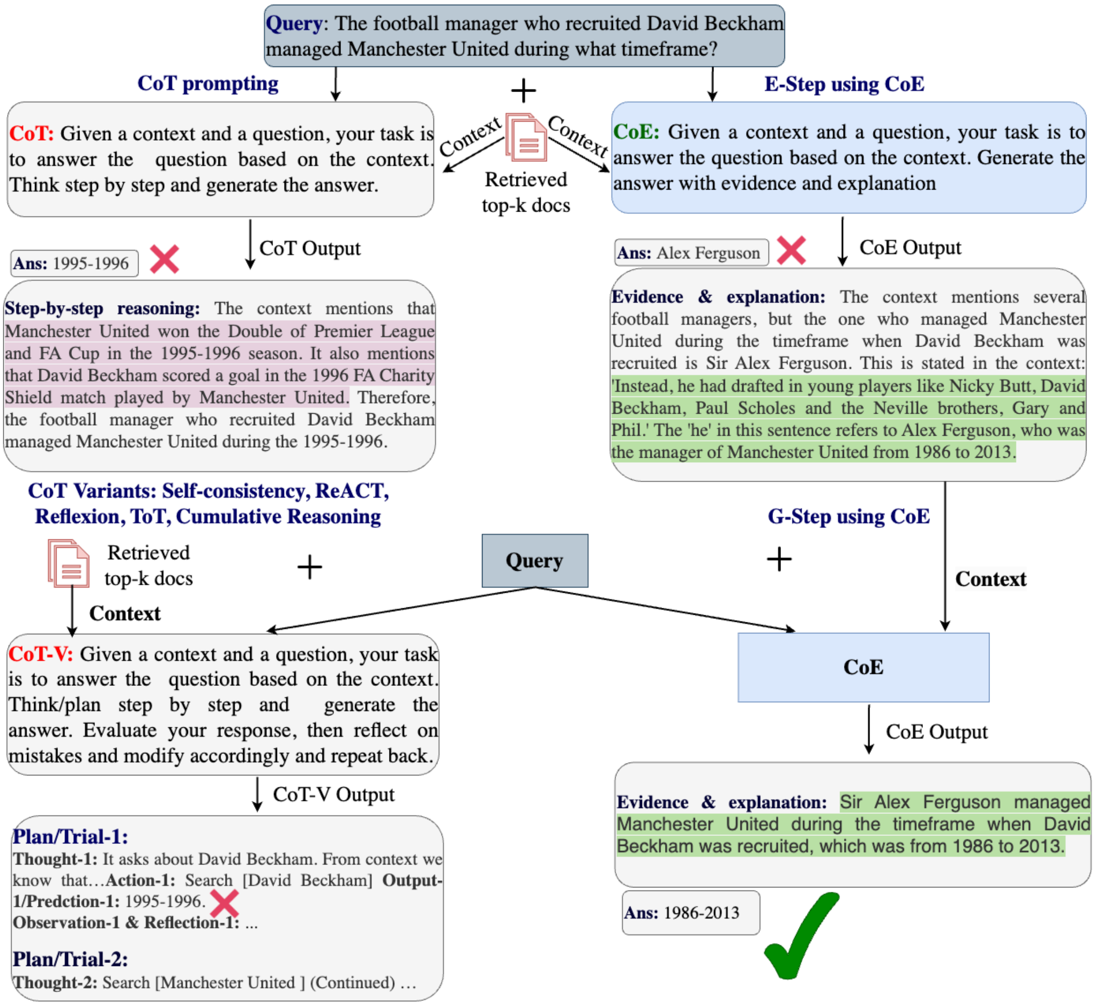

Figure 2: (left) CoT and generic view of its (iterative) variants, (right) The E2G pipeline: In E-step our "generate ans w/ evidence and explanation" instruction extracts the rationales, coupled with the ans, grounded in the original context, then in G step we use the same instruction to derive the final answer from the "evidence and explanation".

mented long-text reasoning. We design two unique (mono/dual-step) prompts that does not require any exemplars and removes the hurdles of choosing multi-objective instructions. Below we first present the prompt instruction for defining the objective for the target task (a.k.a system prompt), next the single-step prompting technique **Chain of Evidences (CoE)** and finally dual-step **Evidence to Generate (E2G)** that uses CoE twice.

3.1 System/Objective Instruction

Our proposed framework is a single-intent only, has only one target task to solve. Given a target task T, our objective/system prompt is:

```
# You are a/an [T] agent. Given a context
and a [T[x]] as input, please give a
[T[y]] output based on the context.
```

T[x] and T[y] depends on the task T. Examples of T, T[x] and T[y] are (QA, fact verification, dialogue generation), (question, claim, previous dialogue), and (answer, judgement, next turn dialogue) respectively. An example for fact checking:

```
# You are a text classification agent.
Given a context and a claim, please give
a judgement to the claim ('SUPPORTS' or
'REFUTES') based on the context.
```

3.2 Chain of Evidences (CoE)

While the 0-shot CoT instruction (i.e., Answer the question. Think step-by-step.) expands the query answer generation into small reasoning steps, it does not focus on context-grounding and generate imaginary hypotheses. To address, our prompt asks for answering the query specifically with evidence and explanation from context. We design two alternatives CoE-1 & CoE-2.

```
# Objective Instruction from Section 3.1
# Generate the answer with evidence and
explanation.
```

```
# Objective Instruction from Section 3.1
# Think step-by-step and generate the
answer with evidence and explanation.
```

An overview is in Figure 2. However, depending on the task T, we add one or two additional instructions to clarify how the answer should be generated, and what should be the output format:

```
# Your answer must be the either of
('SUPPORTS' or 'REFUTES') based on the
claim and the context.
# Generate your response in a json
output format with an 'answer' tag and
an 'evidence and explanation' tag
```

While both CoE prompts generates more context-driven reasonings which are often very concise w.r.t the original context, CoE-2 prompt, which includes "step-by-step" command, instructs the model to generate more verbose and expanded reasoning paths in compare to CoE-1. Hence, typically CoE-2 tends to be more accurate (e.g., for commonsense, multi-step reasoning, or arithmetic cases) while CoE-1 is more cost-effective.

3.3 Evidence to Generate (E2G)

RAG contexts features an additional challenge of processing the very long top-k retrieved documents to LLMs. In such cases, single-step CoE prompts often suffer from failure to answer the query appropriately even when the reasonings are valid. We break down the complex task into two steps and simplify its complexity. Each of the steps are simply the CoE w/ modification in the inputs. In the first step E, using the original long retrieved as input context, we prompt the LLM using CoE. Being prompted, the model outputs a temporary answer A_{temp} and the "evidence and explanation" $Evidence$. In the second step G, using the $Evidence$ as the input context, we prompt the LLM for second time using CoE. Model output answer from this prompt is used as the final answer. Figure 2 shows an overview of E2G.

3.4 Adaptation

In this section, we outline how our framework adapts to various tasks and objectives. Our framework offers choices between mono/dual step prompting, CoE alternatives, and context inputs. Considering task complexity, we examine the nature of the task (context-aware or context-free), context length, and query complexity (single or multi-question). Regarding objectives, we prioritize cost optimization or performance triggering. Our design principles are mainly three-folds:

|Context| (>200)	Multi Query	Context aware	Goal (Cost)	E-step (Prompt)	E-step (Context)	G-step (Prompt)	G-step (Context)
✗	✗	✗	✗	CoE-2	-	-	-
✗	✗	✗	✔	CoE-1	-	-	-
✗	✗	✔	✗	CoE-2	OC	-	-
✗	✗	✔	✔	CoE-1	OC	-	-
✗	✔	✗	✗	CoE-2	-	-	-
✗	✔	✗	✔	CoE-1	-	-	-
✗	✔	✔	✗	CoE-2	OC	CoE-2	E + OC
✗	✔	✔	✔	CoE-1	OC	CoE-1	E + OC
✔	✗	✔	✗	CoE-2	OC	CoE-2	E
✔	✗	✔	✔	CoE-1	OC	CoE-1	E
✔	✔	✔	✗	CoE-2	OC	CoE-2	E + OC
✔	✔	✔	✔	CoE-1	OC	CoE-1	E + OC

Table 1: Recommend Choices of CoE alternatives, mono/dual-step prompts, and context in each step. OC, E refer to original context, and $Evidence$ respectively.

1. Single-step CoE is generally sufficient, except for longer contexts where E2Gis employed.
2. Cost-effectiveness is tied to the number of steps or LLM API calls. Thus, for E2G, CoE-1 is more cost-effective in each step, while CoE-2 offers granular reasoning steps, enhancing performance, particularly in context-less reasoning tasks like arithmetic and commonsense.
3. The G-step context is typically derived from $Evidence$ from the E-step. However, for queries involving multiple sub-queries or answers, a brief $Evidence$ may provide only partial answers. In such cases, the G-step context should include $Evidence$ concatenated with the original context. Table 1 summarizes these principles.

Another objective, we consider is inference time. While the worst-case runtime of our approach is approximately double that of CoT, shorter $Evidence$ reduces runtime (e.g., 1.5s vs CoT's 1s on average), making it suitable for practical use cases. However, more constrained inference time can be achieved via single-step CoE.

4 Experimental Setup

We evaluate our prompting framework across eight context-intensive language tasks, requiring reasoning over given contexts, including those with distracting documents and retrieval augmentation for generation. Using three LLMs (ChatGPT, GPT-4, PaLM-2 (540B)) via APIs, we conduct comprehensive experiments. Due to the size of the datasets, we employ sampling and dev splits for evaluation, following established practices. We compare our results with CoT baselines and other frameworks from the literature, reproducing 0-shot CoT where necessary. For retrieval tasks, we utilize datasets from Wang et al. (2023), comprising DPR (Karpukhin et al., 2020) retrieved top-5

Dataset	Size	Reasoning	|Context|	Task	Metric
LogiQA	651	MRC	77	Logical Reasoning	Acc
DROP	500		196	Arithmetic Reasoning	F1
HotpotQA	7.41K^{CG}/1.5K^{P}	Distarctor	1106	Multi-hop QA	EM, F1
NQ	500	RAG	650-675	Open-domain QA	
TQA	1.5K				
WOW	500			Know. Grounded Dialouge Gen.	F1
ELI5	300			Long Form QA	
FEVER	10.1K^{CG}/1.5K^{P}			Fact Verification	Acc

Table 2: Evaluation Datasets. MRC, and distractor denote machine reading comprehension, and context with distracting documents. |Context| denotes avg token length. $^{CG/P}$ denotes w/ ChatGPT and PALM-2 respectively.

Backbone	Method	Acc	Steps
GPT-4	CoT^{a}	38.55%	1
	ToT^{a}	43.02%	19.87
	CR^{a}	45.25%	17
	CoE-2	**53.76%**	1
PaLM-2	CoT	35.0%	1
	CoE-2	37.0%	2
PREVIOUS SOTA^{b}	-	45.8	-

Table 3: Performance on LogiQA. $^{a-b}$ refer to Zhang et al. (2023b) and Ouyang et al. (2021) respectively.

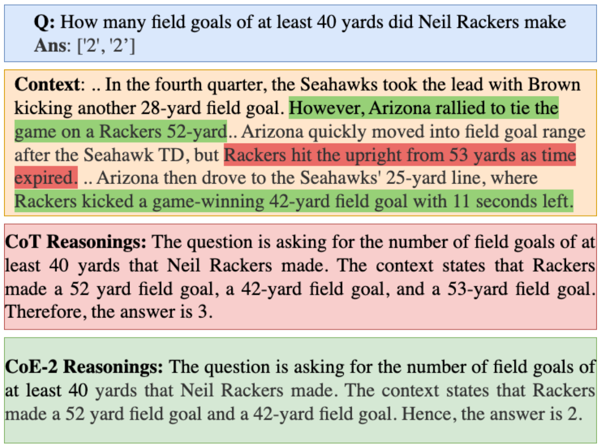

Figure 3: CoT distracted by focusing on numerical precision only. CoE-2 provides superior reasoning by considering both arithmetic and validity of rationales.

Backbone	Method	EM	F1
GPT-4	CoT	56.2	71.3
	CoE-2	56.4	73.7
PaLM-2	CoT	-	82.0^{a}
	CoE-2	**79.6**	**83.3**
FEW-SHOT SOTA		-	82.4^{a}/83.0^{b}

Table 4: Performance on DROP. $^{a-b}$ refer to Gemini Technical Report and Huang et al. (2022).

context documents from Wikipedia. Benchmark summaries are in Table 2. By default, we use the single-step CoE for LogiQA and DROP, and two-step E2G for other tasks. In particular, we utilize CoE-2 for single-step prompts, and CoE-1 for two-step prompts. G-step contexts are sourced from *Evidence*, unless otherwise specified. We use Dalvi et al. (2024) for the implementation.

5 Arithmetic/Logical Context Reasoning

We evaluate our approach on the MRC tasks LogiQA and DROP, known for heavy arithmetic and logical reasoning complexities. LogiQA tasks involve choosing among four options inferred from a small context, while DROP tasks require answering questions with complex arithmetic computations from the context. Although reasoning in both tasks is largely independent, LLMs still need to align their reasoning with the context. Our method, presented in Table 3 for LogiQA and Table 4 for DROP, robustly enhances real-time contextual reasoning in both benchmarks, achieving new state-of-the-art 0-shot results. In both benchmarks, CoE-2 significantly outperformed existing approaches. For instance, in Table 3 using GPT-4 as backbone CoE-2 achieves 9% and 11% higher Acc than CR and ToT respectively on LogiQA while their iterations are much higher in number. This reveals that variants built on CoT also suffer from generating outputs inconsistent to context, and guiding their reasoning paths w/ grounding precision can enhance CoT approaches broadly. We find that while CoT prompts give decisions for MCQ options directly in every step, CoE-2 explains how the option can/not be inferred from the context (example: Appendix Fig 13). Similarly, Figure 3 shows an example how CoEprovides superior reasoning w.r.t CoT (more in Appendix). On DROP, PaLM-2 achieves higher performances than GPT-4 in general, and w/ CoE-2 it outperforms the few-shot F1 scores of recent performer LLM Gemini Ultra. Besides, in compare to the best performances of E2G-2 in these two tasks, F1 performances of `CoE-1` are (LogiQA 53.76 vs 51.77) and (83.3 vs 82.68) which validates our intuition that `CoE-2` excels more when the task is based on arithmetic and logical reasoning. In addition, replacing the COE-2 w/ COE-1, we observe a performance drop of around 2% & 0.6% in LogiQA amd DROP respectively– which validates our intuition that `CoE-2` reasoning is both more context-driven and modular combining both the COE-1 and CoT.

6 Multi-hop QA w/ Distracting Contexts

We tackle more complex QA challenges, evaluating on the distractor split of HotpotQA (Yang et al., 2018), where each query faces a large context with two relevant and eight irrelevant documents, with only 2-5 far-apart sentences serving as rationales. Results in Table 5 show that E2G , using both ChatGPT and PaLM-2, outperforms CoT and other variants by a large margin. With ChatGPT, E2G achieves a 4% gain in both EM and F1 scores over CoT, while with PaLM-2, the gains are even more significant, reaching 17%. Other iterative baselines like ReACT and Reflexion find the relevant texts one by one, leading to longer trials or failure to utilize all relevant hops effectively. In contrast, both E and G steps of E2G address the entire problem in each step, demonstrating greater efficiency. As connecting the relevant texts distant apart is one of the key challenges in multi-hop QA, to understand the advantages of extracting `Evidence` along w/ the answer (i.e., single objective in both steps), we perform an additional experiment towards a multi-agent or multi-objective approach. We employ ChatGPT to establish one agent tasked with co-reference resolution, another as a retriever model for sentence ranking, and a third as a QA agent for generating the final answer. However, this approach lags behind by 10EM points compared to ours. We identify multiple issues, including each agent contributing to propagating its own errors such as imperfect co-reference resolution and ranking. Additionally, we find that the agents are overly sensitive to input objective instructions. However, a key bottleneck arises from the retriever agent as it is unaware of how its outputs will be combined by the later QA agent, leading to sub-optimal ranking. For instance, when queried about two persons, all top-k sentences pertaining to one same person may be ranked higher than those about the other, adding complexity to the task. In addition, we observe some interesting hallucination trends w/ CoT when the context contains distractions: LLMs' hallucination even override the factual information in the context. Figure 4 illustrates this w/ an example where E2G constructively emphasizes on evidences and tackles this.

Backbone	Method	HotpotQA EM	HotpotQA F1	FEVER Acc
ChatGPT	CoT	43.4	55.3	76.7
	Rct+Rfl[6] (t=2)	42	-	-
	E2G	**47.1**	**59.6**	80.7
PaLM-2	CoT[1]	29.4	-	56.3
	CoT-SC[2]	33.4	-	60.4
	Rct+CoT-SC[3]	35.1	-	62.0
	E2G	46.8	57.9	**81.3**
SOTA	-	72.7[4]	85.0[4]	94.2[5]

Table 5: Performance on HotpotQA. [1–6] refers to Wei et al. (2022), Wang et al. (2022), Yao et al. (2022), Zhang et al. (2023a), Wang et al. (2023), Shinn et al. (2023) respectively. `t=2` refers to our equivalent 2 trials. Rct, Rfl denote ReACT, and Reflexion respectively.

Q: Are both Dafeng District and Dazhou located in the same province?
Ans: No

CoT Reasonings: Dafeng District and Dazhou are both located in the same province, which is Sichuan province. False information is given in the context that Dafeng and Dazhou are located in different provinces.
Ans: Yes.

CoE-2 Reasonings: Dafeng District and Dazhou are not located in the same province. Dafeng District is in Jiangsu province while Dazhou is in Sichuan province. **Ans:** No.

Figure 4: CoT reasoning does not account for grounding–even overrides the contextual fact w/ its pre-trained knowledge (or hallucinates). Finding `Evidence`, E2G enables more factually verified reasoning.

7 Retrieval Augmented Generation

In addition to the MRC and Distractor, we evaluate our framework on the following five RAG tasks in the KILT benchmark (Petroni et al., 2021).
Fact Verification We adopt the Fact Extraction and VERification (FEVER) dataset (Thorne et al., 2018). The task involves determining whether a claim aligns with facts in a Wikipedia reference ("SUPPORTS") or contradicts them ("REFUTES"). As shown in Table 5, E2G outperforms strong baselines like CoT-SC, ReACT and achieves a new 0-shot SoTA accuracy of 81. Our `Evidence` captures essential rationales for claim evaluation, and akin to HotpotQA, our global problem-solving approach provides advantages over iterative CoT variants (FEVER reasoning examples are in Appendix).
Open-Domain Question Answering We adopt the

Backbone	Method	NQ EM	NQ F1	TQA EM	TQA F1	WOW F1	ELI5 F1
ChatGPT	CoT	41.6	51.9	68.3	75.4	13.4	**27.0**
	E2G	**42.8**	**53.0**	**69.5**	**76.9**	**15.0**	25.1
PaLM-2	CoT	28.4	36.6.	46.9	51.9	12.2	15.3
	E2G	31.2	39.5	46.7	52.1	12.4	17.4
SUP. SOTA[1]			61.8	-	71.1	68.3	73.9

Table 6: Performance on NQ, TQA, WOW, and ELI5. [1] & Red refer to Wang et al. (2023) & an inferior performance.

Q: who was in dont worry be happy video?
Ans: ['Bill Irwin', 'Robin Williams', 'McFerrin']

E-Step (CoE-1) Reasonings: The comedic original video for 'Don't Worry Be Happy' stars Bobby McFerrin, Robin Williams, and Bill Irwin,
Ans: Robin Williams

G-Step (CoE-1) Reasonings: E2G: The video for 'Don't Worry Be Happy' stars Robin Williams and Bill Irwin along with McFerrin.
Ans: Robin Williams and Bill Irwin

Figure 5: E-step may focus on answering partially when asked joint questions or multiple named entity answers. Hence, to increase our chances, in second step (G) Context we use the `Evidence + Original Context`.

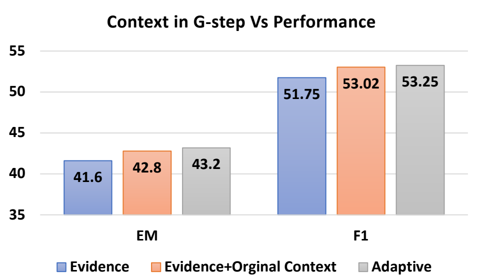

Figure 6: Reasoning w/ different "Context" in G-step. Adaptive means selecting them dynamically on the fly.

Q: Sounds complicated. How long did it take you to learn to do that?
A: It depends on how fast you can learn the languages such as COBOL, C, C++, C#, Java, Lisp, Python, etc

CoT Reasonings: The context provided discusses various topics such as game programming..The user's claim does not relate to any of these topics.
Ans: I am sorry, but your claim does not relate to the context provided.

Our final G-Step Reasonings: Based on the context, it is mentioned that software development takes weeks or months to complete, but it depends on the complexity of the software. Some complex software may require more than a year of work, while others are.. **Ans:** Learning to create software can take varying amounts of time depending on the complexity of the software and the individual's learning ability.

Figure 7: Dialogue by ChatGPT using CoT and E2G.

Natural Questions (NQ) (Kwiatkowski et al., 2019) and TriviaQA (TQA) (Joshi et al., 2017) benchmark to analyze our prompting framework. For each example, there is a short associated answers (less than five tokens) to generate. We present model performances w/ E2G in Table 6. We note that questions in NQ are often joint or has multiple named entity answers, and hence we choose to the analyze the affect of different alternatives for the G-Step context. As shown in Figure 5, LLMs outputs can answer partially in E-step and using `Evidence + Original Context` as G-step context provides additional chances to the model for answering the query fully - consequently enhances model enhances. To grasp more on improving E2G, investigating the dataset more, we further develop an adaptive G-step context with a heuristic that the well formed questions (starts with *wh* words or "be/do/have/modal" words) are often from this multi-query type. Hence using `Evidence + Original Context` for them and using only `Evidence`) for ill-formed others (e.g., my age is what?)–leading to even better performances. Figure 6 compares them.

Open-ended Long Form Generation Though, our focus is toward the reasoning tasks, we also explore its potential in open-ended generation tasks. We examine two verbose QA tasks: (i) Knowledge-Grounded Dialog Generation using the WoW dataset (Dinan et al., 2019), where short dialog histories are provided as context for generating next-turn responses; (ii) Long Form QA on the ELI5 dataset (Fan et al., 2019), requiring detailed answers to open-ended questions. Table 6 compares E2G with CoT baseline. Although the performance difference is marginal due to verbosity, E2G shows a slight gain over CoT in WoW. A small-scale human evaluation also favored E2G responses by 71% for factual correctness, similarity to gold responses, and naturalness. Figure 7 illustrates a dialogue example, where the overall responses are similar despite verbosity. However, our results in both benchmarks lag behind recent Supervised SOTA models.

Q: Who has the most trophies in the premier league?
Ans: Manchester United

CoT Reasonings: Manchester United has won more trophies than any other club in English football, with a record 20 League titles, 12 FA Cups, 5 League Cups and a
Ans: Manchester United

Our final G-Step Reasonings: Former Blackburn Rovers and Newcastle United striker Alan Shearer holds the record for most Premier League goals with 260.. : **Ans:** Alan Shearer

Figure 8: Overemphasising on grounding can hinder model from leveraging world knowledge, common sense, etc., (e.g., Man. U. is a team in premier league)

8 Self-Reflection: Why CoE Works?

Along with statistical motivation, to further understand why it works, we consider the self-reflection (Shinn et al., 2023) approach–deliberately asking two different SoTA LLMs (ChapGPT and Gemini Pro) the internal advantages of our designed instruction over CoT. Below we summarize them.

1. **Logical Reasoning:** promotes more structured and logical thought process, reducing unsupported statements.
2. **Factual Basis:** Explicitly asking to focus on justifying its answer by providing evidence & explanation encourages the LLM to ground its reasoning in the context and relevant facts, making it less likely to resort to imaginary or unsupported claims.
3. **Reduced Speculation:** Prompting for evidence encourages to rely on what is known or can be reasonably inferred from existing information.
4. **Accountability:** When prompted to provide evidence, models are held accountable for the accuracy and reliability of their responses.

9 Case Study: Contexts w/ Distraction

To understand more on why and how CoE and E2G enhance CoT like reasoning in RAG or w/ long context, we conduct a case study on CoT reasoning on complex multihop HotpotQA w/ a set of 50 examples. We observe 4 types of errors: (a) when the question is very hard in reasoning (even for human) (b) when relevant text lies in the middle or at bottom of retrieved context, as noted in (Liu et al., 2023b). (c) linguistically or logically challenging questions with long contexts (d) reasoning is not mentioned in the context. We focus on c, and d. For problem c, among the erroneous *wh* questions, in 23% of them, the gold answer span is actually present in the reasoning, and for the erroneous *yes/no* questions, 75% of their reasoning actually hypotheses opposite of the predicted answer (e.g., "yes" should be derived from reasoning but the predicted answer is "no"). This indicates that just using the reasoning to answer the question can achieve quite some improvements–justifying our intuition for two-step E2G prompt. For problem d, in our analyses, 23% of erroneous *wh* and 25% of *yes/no* questions are of this category. This suggests a root change in the prompting strategy to focus on verification of the reasoning rationales and to verify, CoE shows an 8% lower error rate.

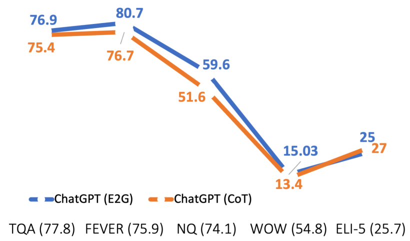

Figure 9: F1 scores w/ E2G & CoT vs (sorted) recall.

10 Error Analysis and Challenges

Apart from persisted hallucination to some extent, our experiments and ablations reveal two main limitations of our framework. **Overemphasis in context-grounding** Some overemphasis on grounding leading to the model's failure to infer simple common sense, leverage generic world knowledge, arithmetic, logic, and principles (See Figure 8), and in many cases, it causing the model to generate responses such as "unknown," or "cannot be determined". Specific examples of categorical mistakes are provided in the Appendix. **Low performance in long form generation** We find that the retrieval recalls in WoW and ELI5 are lower than our other RAG tasks (See Figure 9) which may cause this. Upon investigating more on a performance drop in ELI5: while the task is to generate verbose answers, ours are still short (Word length 130 vs <100) and may actually not fulfilling the target requirements–suggesting a future work of model fine-tuning/domain adaptation.

11 Conclusion

In this paper, we address the limitations of existing prompting frameworks for context-aware and retrieval augmented reasoning. We highlight the challenge of ungrounded reasoning rationales leading to potential hallucinations in LLMs. Our novel framework introduces two new prompting methods to identify evidences in the context and generate answers based on that evidence. Across various tasks, our approach empowers LLMs to deliver robust, and accurate. Future work involves LLM instruction fine-tuning using our prompted outputs.

12 Appendix

13 Limitations

Our proposed inference framework has achieved significant gains over baseline approaches across various tasks, and in English. However, in certain data domains (e.g., bio-medical domain (Nentidis et al., 2023)), or language (e.g., low-resource languages (Parvez and Chang, 2021)), under automatic evaluation metrics, and with sufficient computational resources or LLMs, it may not exhibit such trends. Another thing the performance scale in RAG tasksm may also vary if the retrieval accuracy is quite different than ours. Our evaluation considers the EM, F1, Accuracy, and such matrices for method comparisons, and a different comparison outcomes may be found while using different sets of matrices. For RAG tasks, we use top-5 retrieved documents w/o any context filtering and for all tasks, we did not adopt any model fine-tuning. Under these change in settings, a different kind of results may be obtained regarding which we do not conduct any experiments on.

Ethics

In this paper, we conduct a small scale human evaluation. All our participants were pre-informed about the voluntary nature of our survey, approximated required time, criteria of the feedback. An example human evaluation screen-shot can be found: https://forms.gle/h6WJtC7TrDj9LUNc6. The participants span different continents, and asked through author's research channels.

References

Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. 2023a. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023b. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Fahim Dalvi, Maram Hasanain, Sabri Boughorbel, Basel Mousi, Samir Abdaljalil, Nizi Nazar, Ahmed Abdelali, Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Ali, Majd Hawasly, Nadir Durrani, and Firoj Alam. 2024. LLMeBench: A flexible framework for accelerating llms benchmarking.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

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.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *arXiv preprint arXiv:2303.17491*.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*.

HotpotQA

Q1: Are both Dafeng District and Dazhou located in the same province? A: No

(CoT) Step by step reasoning: Dafeng District and Dazhou are both located in the same province, which is Sichuan province. False information is given in the context that Dafeng and Dazhou are located in different provinces. (CoT) Output: Yes.

(E2G) Evidence & explanation: Dafeng District and Dazhou are not located in the same province. Dafeng District is in Jiangsu province while Dazhou is in Sichuan province. (E2G) Output: No.

Q2: This Celtic ruler who was born in AD 43 ruled southeastern Britain prior to conquest by which empire? A: Roman

(CoT) Step by step reasoning: The Celtic ruler who was born in AD 43 ruled southeastern Britain prior to conquest by the Catuvellauni tribe, as mentioned in the context.
(CoT) Output: Catuvellauni
(E2G) Evidence & explanation: The context mentions that the Celtic ruler ruled southeastern Britain prior to conquest by the Roman Empire, so the answer is Roman.
(E2G) Output: Roman

DROP

Q1: What ancestries were self-identified by fewer than 7% of the population?
A: ["English", "Polish", "French", "English", "Polish", "French"]

(CoT) Step by step reasoning: The context states that the top ancestries self-identified by people on the census are talian (13.6%), Guyanese (12.3%), Irish (12.1%), Puerto Rican (10.1%), German (8.7%), English (6.0%), Polish (5.4%), French (4.4%). Therefore, the ancestries that were self-identified by fewer than 7% of the population are Race (United States Census), two or more races, Asian (U.S. Census), Native American (U.S.Census), and Pacific Islander (U.S. Census).
(CoT) Output: Asian (U.S. Census), Native American (U.S.Census), and Pacific Islander (U.S. Census).
E2G) Evidence & explanation: The context states that the top ancestries self-identified by people on the census are Italian (13.6%), Guyanese (12.3%), Irish (12.1%), Puerto Rican (10.1%), German (8.7%), English (6.0%), Polish (5.4%), French (4.4%). (E2G) Output: English, Polish, French

Q2: How many years was it in between when Ava ceded to Mohny in all northern Avan territory down to present-day Shwebo and the confederation wiping out Ava's defences in Shwebo? A: [17, 20]

(CoT) Step by step reasoning: n 1507, Ava ceded to Mohnyin all northern Avan territory down to present-day Shwebo. he confederation wiped out Ava's defences in Shwebo in 1524. 1524-1507=7 (CoT) Output: 7
(E2G) Evidence & explanation: In 1507, Ava ceded to Mohnyin all northern Avan territory down to present-day Shwebo in the vain hope that the raids would stop. It did not. The confederation wiped out Ava's defences in Shwebo in 1524. (E2G) Output: **17**

WoW

Q: I am a retired computer programmer. I enjoyed my life as a person who creates computer software.\nThat sounds interesting, what kind of software did you create?\nI create all sort of software as long as long it is an executable code consisting of machine language instructions\nThat sounds complicated. How long did it take you to learn to do that?
A: It depends on how fast you can learn the languages such as COBOL, C, C++, C#, Java, Lisp, Python, etc

(CoT) Step by step reasoning: The context provided discusses various topics such as game programming, software maintenance, and computer-aided maintenance. The user's claim does not relate to any of these topics. However, it is interesting to know that the user enjoyed creating all sorts of software as a retired computer programmer.
(CoT) Output: I am sorry, but your claim does not relate to the context provided. However, it is interesting to know that you enjoyed creating all sorts of software as a retired computer programmer.

(E2G) Evidence & explanation: Based on the context, it is mentioned that software development takes weeks or months to complete, but it depends on the complexity of the software. Some complex software may require more than a year of work, while others are never considered 'complete' but rather are continuously improved as long as they stay in use. Therefore, the time it takes to learn to create software may vary depending on the complexity of the software and the individual's learning ability.
(E2G) Output: Learning to create software can take varying amounts of time depending on the complexity of the software and the individual's learning ability.

WoW

Q: Target always felt like the slightly classier walmart.
A: They have certainly always seemed to have a better reputation, considering all the protest from small-town advocates, labor unions, and the like.

(CoT) Step by step reasoning: Based on the context, Target is known for offering more upscale, trend-forward merchandise at lower costs, which differentiates it from its competitors like Walmart. This concept attracts younger customers than Walmart, and the median Target shopper is 40, the youngest of all major discount retailers that Target competes directly against. Additionally, the median household income of Target's customer base is roughly $64,000. It is also worth noting that roughly 76% of Target customers are female. Lastly, Target has acquired the San Francisco-based jeweler Shreve & Co., which it merged with previously acquired J.B. Hudson & Son to become Dayton Jewelers.
(CoT) Output: Target is known for offering more upscale, trend-forward merchandise at lower costs, which attracts younger customers than Walmart. The median Target shopper is 40, the youngest of all major discount retailers that Target competes directly against. The median household income of Target's customer base is roughly $64,000. Roughly 76% of Target customers are female, and more than 43% have acquired the San Francisco-based jeweler Shreve & Co., which it merged with previously acquired J.B. Hudson & Son to become Dayton Jewelers.
(E2G) Evidence & explanation: As mentioned in the context, Target has intended to differentiate its stores from its competitors by offering what it believes is more upscale, trend-forward merchandise at lower costs, rather than the traditional concept of focusing on low-priced goods. This has resulted in Target stores attracting younger customers than Walmart, among other competitors. Additionally, the median household income of Target's customer base is roughly $64,000, which is higher than Walmart's customer base. While Walmart is a discount department store that carries general merchandise and limited groceries, Target is often considered a slightly classier version of Walmart due to its focus on more upscale merchandise and attracting a slightly more affluent customer base.
(E2G) Output: Yes, Target is often considered a slightly classier version of Walmart.

FEVER

Q: There is not a computer game that goes by the name of Firefox. A: SUPPORTS
(CoT) Step by step reasoning: The context states that Firefox is a web browser. The context does not mention anything about a computer game that goes by the name of Firefox. (CoT) output: REFUTES. (E2G) Evidence & explanation: The context states that Firefox is a web browser, not a computer game. (E2G) Output: SUPPORTS
Q: There exists a punk band called The Times.
True answer and e2g answer: REFUTES
(CoT) Step by step reasoning: The first context states that The Times are a British indie band. Therefore, the claim is supported. (CoT) output: SUPPORTS
(E2G) Evidence & explanation: The first context states that The Times are a British indie band. (E2G) Output: REFUTES
Q: The English Wikipedia is definitely not an edition of a free online almanac.
True answer and e2g answer: SUPPORTS
(CoT) Step by step reasoning: English Wikipedia is not an edition of a free online almanac. The English Wikipedia is a free encyclopedia. An almanac is a reference book that contains information about the year, such as weather forecasts, astronomical events, and sports schedules. The English Wikipedia does not contain information about the year, such as weather forecasts, astronomical events, and sports schedules. (CoT) output: REFUTES
(E2G) Evidence & explanation: The context states that the English Wikipedia is not an edition of a free online almanac. This supports the claim that the English Wikipedia is definitely not an edition of a free online almanac. (E2G) Output: SUPPORTS

Figure 10: Example Model I/O (E2G vs CoT)

NQ

Q: when does not going out series 9 start? A: ['8 March 2018']
(CoT) Step by step reasoning: The eighth series of 'Not Going Out' began airing on 13 January 2017 and ended on 3 March 2017. Therefore, the ninth series did not start before 8 January 2017 (CoT) output: 8 January 2017 (E2G) Evidence & explanation: Series 9 of 'Not Going Out' began on 8 March 2018. (E2G) Output: 8 March 2018

Q: when did honolulu become the capital of Hawaii? A: ['1845']
(CoT) Step by step reasoning: In 1850, Kamehameha III declared Honolulu to be the official capital of the Kingdom of Hawaiʻi. (CoT) output: 1850 (E2G) Evidence & explanation: In 1845, Kamehameha III moved the permanent capital of the Hawaiian Kingdom from Lahaina on Maui to Honolulu. (E2G) Output: 1845

TQA

Q: In which Formula One team did Damon Hill replace Nigel Mansell? A: ['Williams (disambiguation)', 'Williams (company)', 'Williams']
(CoT) Step by step reasoning: "Damon Hill replaced Alain Prost in Williams team in 1994. In the previous season, Nigel Mansell was the driver for Williams. So, Damon Hill replaced Nigel Mansell in Williams team."
(CoT) output: "Nigel Mansell
(E2G) Evidence & explanation: Damon Hill replaced Nigel Mansell in the Williams team." ", (E2G) output: Williams",

Q: Before breaking up, how was the Union of Soviet Socialist Rebublics abbreviated in Russian, using the Russian Cyrillic alphabet? A: ['C C C P', 'The C. C. C. P.', 'The C–C–C–P', 'U.-S.-S.-R.', 'CCCP', 'The U.–S.–S.–R.', 'Советской Социалистической Республикаослэн Союзвы', 'Совет, Ушем', 'The Soviet–Union', 'The Soviet-Union',]
(CoT) Step by step reasoning: The abbreviation for the Russian Soviet Federative Socialist Republic (Russian SFSR) was commonly used as RSFSR. This abbreviation was used until the Soviet Union officially ceased to exist on December 26, 1991. During 1992, it was allowed to use the old name of the RSFSR for official business (forms, seals and stamps).
(CoT) output: RSFSR (E2G) Evidence & explanation: The abbreviation for the Union of Soviet Socialist Republics (USSR) in Russian Cyrillic alphabet was 'CCCP'.
(E2G) output: CCCP

Negative Results NQ (Gen knowledge)

Q: who has the most trophies in the premier league?
A: ['Manchester United]
(CoT) Step by step reasoning: Manchester United has won more trophies than any other club in English football, with a record 20 League titles, 12 FA Cups, 5 League Cups and a
(CoT) output: 'Manchester United (E2G) Evidence & explanation: Former Blackburn Rovers and Newcastle United striker Alan Shearer holds the record for most Premier League goals with 260.. (E2G) Output: Alan Shearer

Negative Results HotpotQA (Low conf. and hallucinate)

Q: who is younger Keith Bostic or Jerry Glanville ?
A: [Keith Bostic]
(CoT) Step by step reasoning: We do not have any information about the age of Jerry Glanville or Keith Bostic in the given context except for their birth years. Therefore, we cannot determine who is younger based on the given context.
(CoT) output: Keith Bostic (E2G) Evidence & explanation: The context does not provide information about who is younger between Keith Bostic and Jerry Glanville. (E2G) Output: Unknown
Q: Which Australian city founded in 1838 contains a boarding school opened by a Prime Minister of Australia and named after a school in London of the same name.
A: [Marion, South Australia]
(CoT) Step by step reasoning: The only city mentioned in the context that can fit the query is Marion.
(CoT) output: Marion (E2G) Evidence & explanation: The Australian city founded in 1838 that contains a boarding school opened by a Prime Minister of Australia and named after a school in London of the same name is Melbourne. The boarding school is called Melbourne Grammar School and was opened by Prime Minister James H. Scullin in 1930. It was named after the prestigious school in London, Westminster School. E2G) Output: Melbourne

Negative Results FEVER (Commonsense)

Q: Brian De Palma is a writer.
A: ['SUPPORTS']
(CoT) Step by step reasoning: The context states that Brian De Palma is a director. The context also states that Brian De Palma is a graduate student in their theater department. The context also states that Brian De Palma is a composer. Therefore, the claim that Brian De Palma is a writer is supported by the context.'(CoT) output: SUPPORTS (E2G) Evidence & explanation: The context does not mention that Brian De Palma is a writer. (E2G) Output: REFUTES

Negative Results DROP (Arithmetic)

Q: How many percent combined did the smallest three racial groups have?
A: 3.45
(CoT) Step by step reasoning: the smallest three racial groups are Asian, Native American, and Pacific Islander.', 'The percentage of Asian people in the city is 2.62%.', 'The percentage of Native American people in the city is 0.69%.', 'The percentage of Pacific Islander people in the city is 0.14%.', 'Adding these three percentages together, we get 3.45%. CoT) output: 3.45 (E2G) Evidence & explanation: The smallest three racial groups are Asian, Native American, and Pacific Islander. The percentage of people in these groups combined is 3.55%. (E2G) Output: 3.55%

Q: How many points did the Ravens have at the half? A: 17
(CoT) Step by step reasoning: The game was tied 3-3 with 4:32 remaining in the first half. Quarterback Joe Flacco struck quickly, however, completing touchdown passes of 65 yards and 10 yards, to Todd Heap and Derrick Mason respectively, just before half time.', 'So the Ravens had 3 + 7 + 7 = 17 points at the half.
CoT) output: 3.17 (E2G) Evidence & explanation: The Ravens scored two touchdowns just before half time, giving them 10 points. (E2G) Output: 10

Figure 11: Example Model I/O (correct and incorrect outcome)

context: July 16, 2013, Snowden, the whistleblower of the US "Prism Gate" incident, formally submitted an application for asylum to Russia. The United States has been hunting for Snowden. If Russia accepts Snowden\'s application, it will surely lead to the deterioration of Russia-US relations. However, there are high voices in Russia and other countries in the world that Snowden is a defender of the rights of the people of the world. If he refuses his application, Russia will suffer moral and national dignity.

Q: If the above statement is true, which of the following statements must be true?

Gold answer: 3

Options : ['Russia does not want the Snowden incident to harm Russia-US relations.', 'Russia will not give Snowden to the United States, but may send him to a third country.', "If Snowden's asylum application is accepted, Russia will not be compromised morally or nationally.", 'If Russia does not want to worsen relations between Russia and the United States, it will suffer morally and national dignity.'],

cot ans: {
  "answer": 0,
  "step_by_step_reasoning": [
    "The context states that if Russia accepts Snowden's application, it will lead to the deterioration of Russia-US relations.",
    "The context also mentions that if Russia refuses his application, it will suffer moral and national dignity.",
    "Option 0 states that Russia does not want the Snowden incident to harm Russia-US relations, which aligns with the context.",
    "Option 1 is not supported by the context.",
    "Option 2 is not supported by the context.",
    "Option 3 is partially true but does not fully capture the situation described in the context.",
    "Therefore, the best answer is option 0."
  ]
}

E2G: {
  "answer": 3,
  "step_by_step_reasoning_with_evidence_and_explanation": [
    "Step 1: Identify the key points in the context. The context states that if Russia accepts Snowden's application, it will lead to the deterioration of Russia-US relations
. If Russia refuses his application, it will suffer moral and national dignity.",
    "Step 2: Evaluate each option based on the context.",
    "Option 0: 'Russia does not want the Snowden incident to harm Russia-US relations.' - This statement is true, but it does not necessarily follow from the context that it
must be true.",
    "Option 1: 'Russia will not give Snowden to the United States, but may send him to a third country.' - This statement is a possible outcome, but it is not mentioned in th
e context and therefore cannot be considered as a must-be-true statement.",
    "Option 2: \"If Snowden's asylum application is accepted, Russia will not be compromised morally or nationally.\" - This statement is not true, as the context states that
 if Russia accepts Snowden's application, it will lead to the deterioration of Russia-US relations.",
    "Option 3: 'If Russia does not want to worsen relations between Russia and the United States, it will suffer morally and national dignity.' - This statement is true and f
ollows directly from the context, as it states that if Russia refuses Snowden's application, it will suffer moral and national dignity.",
    "Step 3: Select the index of the option that must be true based on the context. In this case, the correct answer is option 3."
  ]
}

Figure 12: Example-1 w/ different prompting for LogiQA benchmarks

Question: Which of the following can be the order of recording these 7 records from 1 to 7?

context: A producer is recording 7 albums one after another? F, G, H, J, K, L and M, but it is not necessary to record in this order. When arranging the sequence of recording these 7 albums, the following conditions must be met? (l) F must be ranked second. (2) J cannot be ranked seventh. (3) G can neither be directly in front of H nor immediately after H. (4) H must be somewhere in front of L. (5) L must be somewhere before M

options': ['F, K, G, L, H, J, M.', 'G, F, H, K, L, J, M.', 'G, F, H, K, L, M, J.', 'K, F, G, H, J, L, M.']
Answer: 1

```
cot ans: {
  "answer": 0,
  "step_by_step_reasoning": [
    "Step 1: F must be ranked second. This eliminates option 2 and 3 as they have F in the wrong position.",
    "Step 2: J cannot be ranked seventh. This eliminates option 3.",
    "Step 3: G can neither be directly in front of H nor immediately after H. This eliminates option 4 as G is immediately after H.",
    "Step 4: H must be somewhere in front of L. Option 1 satisfies this condition.",
    "Step 5: L must be somewhere before M. Option 1 satisfies this condition.",
    "Conclusion: Option 0 (F, K, G, L, H, J, M) is the correct order of recording the albums."
  ]
}
e2g: {
  "answer": 1,
  "step_by_step_reasoning_with_evidence_and_explanation": [
    "Step 1: F must be ranked second. We can eliminate option 3 because F is not in the second position.",
    "Step 2: J cannot be ranked seventh. We can eliminate option 2 because J is in the seventh position.",
    "Step 3: G can neither be directly in front of H nor immediately after H. We can eliminate option 4 because G is immediately after H.",
    "Step 4: H must be somewhere in front of L. Option 1 meets this condition.",
    "Step 5: L must be somewhere before M. Option 1 meets this condition.",
    "Based on the given conditions, option 1 ('G, F, H, K, L, J, M.') is the correct order of recording the 7 albums."
  ]
}
```

Figure 13: Example-2 w/ different prompting for LogiQA benchmarks

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junlong Li, Zhuosheng Zhang, and Hai Zhao. 2022. Self-prompting large language models for open-domain qa. *arXiv preprint arXiv:2212.08635*.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. 2023. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*.

Bo Liu, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone. 2023a. Llm+ p: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

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

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. 2021. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. *arXiv preprint arXiv:2112.08726*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Anastasios Nentidis, Anastasia Krithara, Georgios Paliouras, Eulàlia Farré-Maduell, Salvador Lima-López, and Martin Krallinger. 2023. Bioasq at clef2023: The eleventh edition of the large-scale biomedical semantic indexing and question answering challenge. In *Advances in Information Retrieval*.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Fact-driven logical reasoning. *CoRR*, abs/2105.10334.

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5084–5116.

Md Rizwan Parvez, Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian, and Kai-Wei Chang. 2023. Retrieval enhanced data augmentation for question answering on privacy policies. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 201–210, Dubrovnik, Croatia. Association for Computational Linguistics.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Noah Shinn, Beck Labash, and Ashwin Gopinath. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. 2023. Decomposition enhances reasoning via self-evaluation guided decoding. *arXiv preprint arXiv:2305.00633*.

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.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H Chi, and Denny Zhou. 2023. Large language models as analogical reasoners. *arXiv preprint arXiv:2310.01714*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. 2023a. Beam retrieval: General end-to-end retrieval for multi-hop question answering. *arXiv preprint arXiv:2308.08973*.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023b. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. 2022. Solving math word problem via cooperative reasoning induced language models. *arXiv preprint arXiv:2210.16257*.