KNOWLEDGE CROSSWORDS: Geometric Knowledge Reasoning with Large Language Models

Wenxuan Ding[*1] **Shangbin Feng**[*2] **Yuhan Liu**[3] **Zhaoxuan Tan**[4]
Vidhisha Balachandran[5] **Tianxing He**[2] **Yulia Tsvetkov**[2]
[1]The Hong Kong University of Science and Technology [2]University of Washington
[3]Xi'an Jiaotong University [4]University of Notre Dame [5]Carnegie Mellon University
wdingaj@connect.ust.hk, shangbin@cs.washington.edu

Abstract

We propose KNOWLEDGE CROSSWORDS, a geometric knowledge reasoning benchmark consisting of incomplete knowledge networks bounded by structured factual constraints, where LLMs are tasked with inferring the missing facts to meet all constraints. The novel setting of geometric knowledge reasoning necessitates new LM abilities beyond existing atomic/linear multi-hop QA, such as backtracking, verifying facts and constraints, reasoning with uncertainty, and more. KNOWLEDGE CROSSWORDS contains 2,101 individual problems, covering diverse knowledge domains, and is further divided into three difficulty levels. We conduct extensive experiments to evaluate existing LLMs and approaches on KNOWLEDGE CROSSWORDS. Results demonstrate that baseline approaches struggle with larger knowledge networks and semantically-equivalent entity distractors. In light of their limitations, we propose two new approaches, STAGED PROMPTING and VERIFY-ALL, to augment LLMs' abilities for error-aware backtracking and constraint verification. Our VERIFY-ALL significantly outperforms prior methods and is more robust towards problems in the hard subset. Further analysis shows that geometric knowledge reasoning poses new challenges to LLMs' knowledge abilities, particularly in robustness towards varying option orders, complex structural constraints in knowledge networks, "none of the above" scenarios, and more.[1]

1 Introduction

Large language models (LLMs) encode wast amounts of world knowledge in model parameters (Petroni et al., 2019; Yu et al., 2023a). Existing tasks and datasets assess LLM knowledge abilities mostly by focusing on atomic (e.g., open-domain QA) (Rajpurkar et al., 2016; Das et al., 2022; Joshi et al., 2017) or linear (e.g., multi-hop QA) (Press et al., 2023; Yang et al., 2018; Ho et al., 2020) settings, extracting one fact or a fixed concatenation of facts from LLMs. However, knowledge (and language) is naturally structured (Reagans and McEvily, 2003), going beyond linear arrangements, involving complex structural attributes, and forming an interweaving network that connects various entities and relations through multiple chains as illustrated in Figure 1.

Consequently, we ask: *Can LLMs extend beyond linear compositionality and aggregate information from multiple chains along with various knowledge constraints?* Specifically, can LLMs, with the help of their internal parametric knowledge and inherent reasoning patterns, infer missing facts in a network? We term such ability *geometric knowledge reasoning*. While compositional QA has been explored in the constrained setting of *external* knowledge bases (Zelle and Mooney, 1996; Cui et al., 2017; Ye et al., 2022; Neelam et al., 2022; Xie et al., 2022), we aim to investigate whether LLMs could reason with non-linear fact networks solely relying on their *internal* parametric knowledge. Formally, we define *geometric knowledge reasoning* as reasoning over a network by inferring missing entities based on the given contextual information, where such networks cannot be simply broken down into chains (like linear reasoning), risking the loss of structural information and constraints. Geometric knowledge reasoning with LLMs naturally necessitates new LLM abilities beyond those encountered in atomic or linear tasks, such as composing knowledge across multiple chains, reasoning with uncertainty, fact verification, error-aware backtracking, and more. Since state-of-the-art LLMs are trained on linear sequences of texts devoid of explicit structure, it is underexplored whether they could effectively apply their linearly acquired knowledge to solve geomet-

[*] equal contribution
[1]Code and data are publicly available at https://github.com/Wenwen-D/KnowledgeCrosswords.

Findings of the Association for Computational Linguistics: ACL 2024, pages 2609–2636
August 11-16, 2024

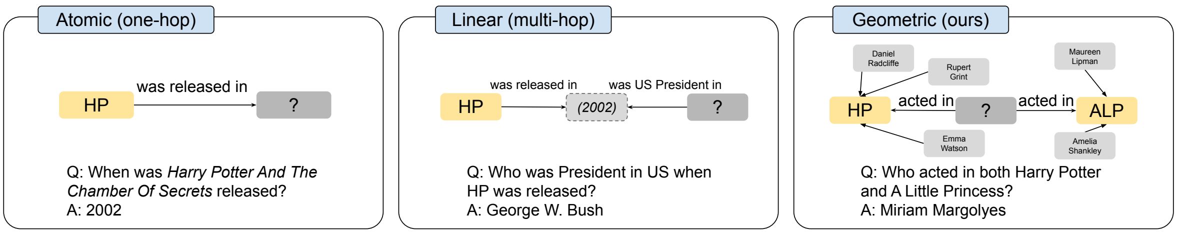

Figure 1: Illustration of the differences of atomic, linear (multi-hop), and *geometric knowledge reasoning*. Each step of atomic or linear QA leads to a unique and definite (intermediate) answer, while multiple candidates in each step should be jointly considered to satisfy structural constraints in geometric knowledge reasoning.

ric reasoning tasks. LLMs with strong geometric knowledge reasoning abilities could serve as versatile relational databases, allowing controllable access to parametric knowledge through SQL-like conditioned prompting.

To this end, we propose KNOWLEDGE CROSSWORDS, a geometric knowledge reasoning dataset with 2,101 problems evaluating to what extent LLMs could achieve such desiderata. Each knowledge crossword consists of a list of constraints representing an incomplete fact network, and LLMs need to reconstruct the knowledge network while ensuring that all factual constraints are met. To solve knowledge crosswords, LLMs should ideally evaluate candidates for each blank, jointly consider factual constraints, verify intermediate solutions, and backtrack when encountering factual errors, until all constraints are met. For each incomplete fact network, we generate three sets of distractors as options that are progressively more plausible, resulting in *easy*, *medium*, and *hard* subsets for fine-grained evaluation. Each problem also comes with two settings: *w/o knowledge*, where LLMs solve knowledge crosswords solely with parametric knowledge; *w/ knowledge*, where a helpful (and noisy) paragraph is prepended to each problem.

We conduct extensive experiments to evaluate LLMs and approaches on KNOWLEDGE CROSSWORDS, ranging from simple zero-shot prompting to advanced ones such as self-consistency (SC) (Wang et al., 2022) and least-to-most prompting (LTM) (Zhou et al., 2022). Results demonstrate that baselines struggle with problems in the hard subset and have significant performance drops in the *w/o knowledge* setting. Advanced prompting methods such as SC and LTM barely improve LLMs due to their reliance on left-to-right reasoning patterns. To address these challenges, we propose two new instruction-based techniques, STAGED PROMPTING and VERIFY-ALL, aiming at augmenting LLMs' abilities for backtracking, constraint verification, and more. STAGED PROMPTING guides LLMs through an elaborate problem-solving process that progressively solves and simplifies the problem blank by blank, while VERIFY-ALL proposes candidates for all blanks and verifies them simultaneously. We find that VERIFY-ALL achieves top performance and is more robust with harder problems, while the success of STAGED PROMPTING is contingent on stronger base LLMs. Further analysis reveals geometric knowledge reasoning poses great challenges to LLM knowledge abilities, as they could be susceptible to a variety of factors such as option order, "none of the above" scenarios, number of distractors, special structural patterns, and more. We envision geometric knowledge reasoning as a challenging research question and KNOWLEDGE CROSSWORDS as a comprehensive testbed to evaluate LLM knowledge abilities in more complex and structured settings.

2 KNOWLEDGE CROSSWORDS

We propose KNOWLEDGE CROSSWORDS, a geometric knowledge reasoning benchmark to evaluate whether LLMs could reason with incomplete fact networks bounded by geometric constraints (Appendix B). An example knowledge crossword is presented in Figure 2.

Definition Each knowledge crossword consists of a question graph $\mathcal{G}_Q = \{(h, r, t) | h, t \in \mathcal{V}_Q, r \in \mathcal{R}\}$, where $\mathcal{V}_Q$ is the set of entities represented as nodes of $\mathcal{G}_Q$ and $\mathcal{R}$ is the set of all possible relations between entities. Each (h, r, t) in $\mathcal{G}_Q$ denotes a directed edge representing a factual association such as *(Marvin Minsky, has won prize, Turing award)*. In the question graph $\mathcal{G}_Q$, certain nodes $b_i \in \mathcal{V}_Q$ are masked out as blanks $\mathcal{B} = [b_1, b_2, \dots, b_{|\mathcal{B}|}]$ for QA. The goal of each knowledge crossword is to find one combination

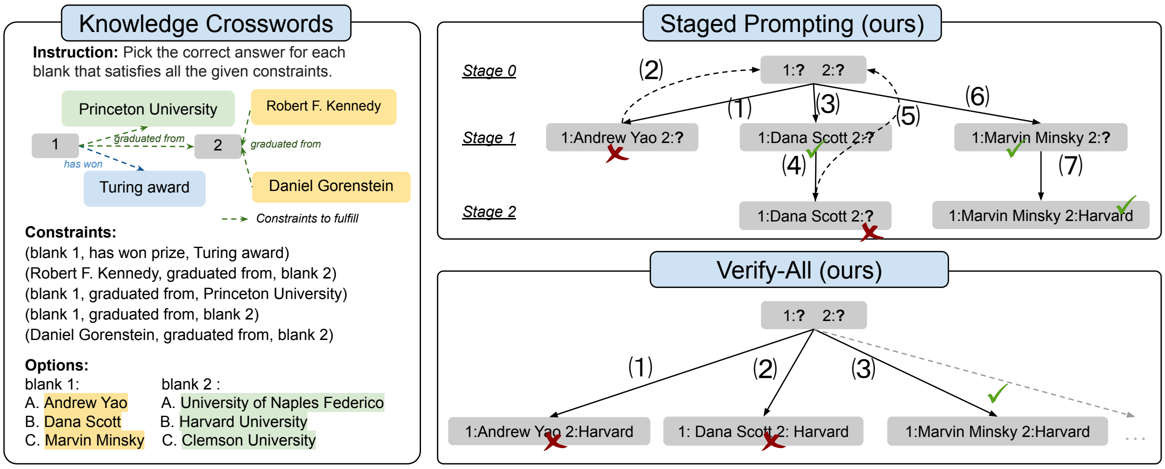

Figure 2: Overview of KNOWLEDGE CROSSWORDS and two proposed approaches, STAGED PROMPTING and VERIFY-ALL. Each knowledge crossword includes task instruction, factual constraints, and multiple-choice QA options. In each stage of STAGED PROMPTING, LLMs ① *solve* one blank based on one remaining constraint; ② *update* the status by filling in the proposed answer; then ③ *verify* filled constraints to proceed or backtrack. In VERIFY-ALL, LLMs propose a combination of ① *candidates* and ② *verify* all constraints with those candidates, and repeat this process until all constraints are met.

of answers for all blanks $\mathcal{A} = [a_1, a_2, \ldots, a_{|\mathcal{B}|}]$ that satisfies all factual associations represented as geometric constraints in the question graph $\mathcal{G}_Q$.

Data Source We resort to encyclopedic knowledge graphs, specifically YAGO (Suchanek et al., 2023), as scaffolds of geometric knowledge reasoning to construct the KNOWLEDGE CROSSWORDS benchmark. Different from existing KBQA datasets (ComplexWebQuestions, Talmor and Berant (2018), GrailQA, Gu et al. (2021), *inter alia*) where LMs are required to reason with *external* KBs, LLMs solve knowledge crosswords with their *internal* parametric knowledge. We conduct preprocessing to remove certain relations in YAGO that are location-related, time-sensitive, or not self-evident. This is to ensure that the KNOWLEDGE CROSSWORDS is minimally affected by question ambiguity (Min et al., 2020; Cole et al., 2023), outdated knowledge (Yu and Ji, 2023; Hernandez et al., 2023), etc. We obtain the filtered knowledge graph as $\mathcal{KG} = \{(h, r, t) | h \in \mathcal{H}, r \in \mathcal{R}, t \in \mathcal{T}\}$, where $\mathcal{H}$, $\mathcal{R}$, and $\mathcal{T}$ are the sets of heads, relations, and tails respectively.

Question Graphs We first adopt two hyperparameters to control the property and difficulty of the generated question graphs (incomplete fact networks): *question graph size* s_G, denoting the total number of nodes in a question graph, and *blank size* s_B, representing the number of nodes masked out as blanks that need to be filled. We start from a random center node c and retrieve the k-hop neighborhood of c as $\mathcal{G}_{\mathcal{N}}^{(c)}$. We then downsample $\mathcal{G}_{\mathcal{N}}^{(c)}$ by randomly removing nodes with degrees higher than a dynamic threshold t_d in $\mathcal{KG}$, until the largest weakly connected component in $\mathcal{G}_{\mathcal{N}}^{(c)}$ has a size no greater than s_G. This is motivated by the fact that entities with higher degrees are presumably less typical and more ambiguous, resulting in problem ambiguity (Shomer et al., 2023; Qian et al., 2023). We refer to the largest connected component in downsampled $\mathcal{G}_{\mathcal{N}}^{(c)}$ as an answer graph $\mathcal{G}_{\mathcal{A}}$.

We then randomly select s_B nodes in $\mathcal{G}_{\mathcal{A}}$ with degrees larger than a threshold t_b in $\mathcal{G}_{\mathcal{A}}$ and mask them out as blanks B to obtain a question graph. These high-degree blanks would be rich in geometric associations and provide more constraints to work with. The question graph is then linearized in triplet format, since converting interconnected triples into plain natural language and vice versa are noisy and prone to introduce errors and biases (Bai et al., 2023; Min et al., 2023). By employing hyperparameters and thresholds such as s_G and s_B, the dataset comes with built-in difficulty control measures to controllably generate diversified problems. As the question graph generation step does not guarantee answer uniqueness, we exhaustively search answers for each $\mathcal{G}_Q$ in $\mathcal{KG}$ and only retain those with one valid answer combination.

Negative Sampling We mainly consider KNOWLEDGE CROSSWORDS in a multiple-choice setting, where several candidates are provided for each blank in $\mathcal{G}_Q$. This would require negative sampling for each blank to provide distractors in addition to the correct answer, while we identify a taxonomy of three progressive rules for distractors from loose to strict:

- Rule #1: **Semantic Role**: If a blank b_i is the head (or tail) of an edge with relation r_i, then the distractor for b_i should be selected from the set of heads (or tails) of edges with the same r_i.
- Rule #2: **Network Proximity**: The distractor for b_i should occur in the neighborhood $\mathcal{G}_{\mathcal{N}}^{(c)}$ around c, which further ensures that distractors are likely to be in a similar context as b_i.
- Rule #3: **Definite Constraint**: If the other end of the edge that the blank b_i is incident to is known, then we say such edge is a definite constraint for b_i, and the distractor should satisfy at least one of such definite constraints for b_i to fulfill Rule #3. Such distractors impose higher demands on LMs in the sense that LMs should jointly consider all constraints to exclude these distractors.

As a result, we obtain three negative sampling strategies with varying difficulty implications for knowledge crosswords: *easy*, where distractors only meet Rule #1; *medium*, where distractors meet Rule #1 and #2; *hard*, where distractors meet Rule #1, #2, and #3. We opt to separately assign multiple options for each blank in $\mathcal{G}_Q$, using either *easy*, *medium*, or *hard* strategies, resulting in three subsets of knowledge crosswords with increasing difficulty levels.

Relevant Knowledge Each knowledge crossword comes with two settings: W/O KNOWLEDGE, where LLMs solely rely on internal parametric knowledge to solve knowledge crosswords; W/ KNOWLEDGE where a helpful but noisy passage is prepended to the problem description. These knowledge passages contain both helpful information about the correct answers and irrelevant information generated by the three proposed negative sampling rules. Useful and irrelevant knowledge is then sampled to 1:3, shuffled, and presented before each knowledge crossword. LLMs would need to dynamically select relevant and useful information to facilitate geometric knowledge reasoning, which poses new challenges to LLMs (Shi et al., 2023a).

Subset	#Qs	Avg. #Nodes	Avg. #Edges	Avg. #Blanks
EASY	869	7.28	6.63	2.89
MEDIUM	873	7.28	6.64	2.89
HARD	359	7.09	6.41	2.62

Table 1: Statistics of the KNOWLEDGE CROSSWORDS Benchmark. We report the number of questions and the average number of nodes, edges, and blanks for each subset with different negative sampling difficulty.

Evaluation Metrics We evaluate performance on KNOWLEDGE CROSSWORDS with two metrics: *Partial-Credit* (PC), indicating the portion of blanks that have been answered correctly; *Full-Credit* (FC), indicating whether all blanks are correct in a given knowledge crossword. Formally,

$$\text{PC} = \frac{\sum_{i=1}^{s_B} \mathbb{1}[a'_i = a_i]}{s_B}, \quad \text{FC} = \mathbb{1}[\text{PC} = 1]$$

where a'_i denotes the prediction of blank b_i by LLMs and $\mathbb{1}[\cdot]$ denotes the indicator function.

Benchmark Statistics We obtain 873 valid question graphs with different levels of scales and sparsity. Each question graph is then used to construct three problems using the three levels of negative sampling difficulty, *easy*, *medium*, and *hard*, resulting in a total of 2,101 problems and enabling fine-grained evaluation. The problems are described in English and the benchmark statistics are shown in Table 1.

3 Experiment Settings

3.1 Baselines

We evaluate various prompting approaches on KNOWLEDGE CROSSWORDS, including Zero-Shot (ZERO-SHOT) prompting, Few-Shot in-context learning (FEW-SHOT), Chain-of-Thought prompting (COT), CoT with Self-Consistency (COT+SC), and Least-to-Most prompting (LTM). Besides, we adopt the RANDOM baseline which refers to randomly selecting an option for each blank. We also present an UPPERBOUND baseline, where we present oracle knowledge to the LLM, *i.e.* the constraints in G_Q filled with correct answers.

3.2 Models and Settings

Unless otherwise specified, we use ChatGPT (GPT-3.5-TURBO) as the base language model in our experiments, and we additionally test out GPT-4 and open-source models such as Llama 2 (Touvron et al., 2023). For Few-Shot prompting techniques

Method	W/ KNOWLEDGE						W/O KNOWLEDGE					
	easy		**medium**		**hard**		**easy**		**medium**		**hard**	
	PC	FC	PC	FC	PC	FC	PC	FC	PC	FC	PC	FC
RANDOM	34.3	6.1	34.2	5.5	33.5	8.4	34.3	6.1	34.2	5.5	33.5	8.4
UPPERBOUND	98.8	96.7	99.1	97.4	91.8	82.2	-	-	-	-	-	-
ZERO-SHOT	97.3	93.7	97.4	94.2	77.9	55.4	81.3	57.1	83.3	60.6	57.2	24.8
FEW-SHOT	97.8	93.2	97.6	93.5	78.8	54.0	83.7	60.8	84.7	63.3	56.8	25.3
COT	94.6	86.5	95.0	88.9	77.9	56.3	74.0	44.0	76.4	48.5	55.7	27.0
COT+SC	95.9	89.8	96.6	91.2	78.7	57.4	75.2	45.8	77.3	49.1	56.7	28.4
LTM	86.0	68.9	86.3	68.6	69.6	43.5	75.6	47.3	76.6	48.2	51.1	19.2
STAGED PROMPTING	91.9	81.6	91.2	80.4	70.5	44.5	64.3	34.3	67.4	38.3	47.9	15.8
VERIFY-ALL	**98.6**	**96.1**	**98.7**	**96.2**	**83.9**	**64.6**	**84.5**	**62.3**	**86.1**	**66.9**	**59.7**	**29.8**
Experiments with GPT-4												
STAGED PROMPTING	**99.1**	**98.8**	**96.3**	95.6	**95.4**	**94.2**	75.4	70.7	78.8	74.0	52.3	32.4
VERIFY-ALL	98.1	98.1	95.7	**95.7**	92.8	90.5	**88.0**	**83.4**	**89.5**	**85.5**	**59.5**	**38.6**

Table 2: FC and PC with GPT-3.5-TURBO unless otherwise specified. The best results are **bold-faced**, and the second-best ones are underlined. Notably, VERIFY-ALL outperforms the second-best baselines by 7.2% and 1.4% (FC) on the hard subset under W/ KNOWLEDGE and W/O KNOWLEDGE respectively.

(FEW-SHOT, COT, COT+SC, LTM), we present 5 in-context exemplars. The sampling temperature τ is set to 0.1 except for COT+SC; we sample 5 Chain-of-Thought responses with temperature $\tau = 0.7$ for the CoT with Self-Consistency baseline.

4 Our Approach

We hypothesize that the left-to-right reasoning patterns in autoregressive language models (Yao et al., 2023) and prompting approaches (discussed in section 3.1) would fall short of solving knowledge crosswords, which require backtracking, maintaining problem states, verifying existing information, reasoning with structured constraints, and more. To this end, we introduce two instruction-based methods that promote these abilities, illustrated with a detailed example in Figure 2.

4.1 STAGED PROMPTING

The STAGED PROMPTING approach divides geometric knowledge reasoning into stages involving three steps: *solve*, *update*, and *verify*. At the beginning of each stage, LLMs maintain a current status of solved blanks and unresolved constraints (edges that involve unsolved blanks). In the *solve* step, LLMs propose a candidate for an unsolved blank based on internal knowledge; in the *update* step, LLMs update unsolved constraints using the newly proposed candidate for an associated blank; in the *verify* step, LLMs reflect on the updated constraints in the *update* step and judge their validity. If an invalid factual association is identified as a result of the proposed candidate, LLMs backtrack to the problem status in the previous stage and propose another option; otherwise, LLMs proceed to tackle the remaining blanks until all blanks are filled and all constraints are met.

4.2 VERIFY-ALL

While STAGED PROMPTING presents an elaborate problem-solving process that tackles challenges such as backtracking and status updates, such complex reasoning might be hard to learn in context for LLMs. We additionally propose the VERIFY-ALL approach: candidates for each blank are simultaneously proposed, rather than in separate stages. A verification step is then employed to assess the validity of all filled constraints using these proposed candidates. If an error is detected, the LM should backtrack and propose an alternative set of candidates until no error is found.

5 Results

We evaluate approaches on KNOWLEDGE CROSSWORDS and present results in Table 2.

LLMs have preliminary abilities for geometric knowledge reasoning. Table 2 shows that all investigated approaches outperform the RANDOM

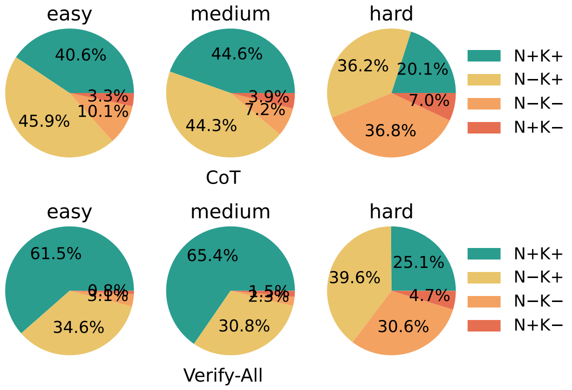

Figure 3: Problem distribution based on the correctness under the W/ KNOWLEDGE and W/O KNOWLEDGE settings using COT and VERIFY-ALL. The results indicate that while easier problems are mainly hindered by a lack of knowledge, the bottleneck for hard problems lies in geometric knowledge reasoning abilities.

baseline, while LLMs could achieve 90+ FC scores on the *easy* subset and *w/ knowledge* setting. However, model performance (FC) drops by 29.7% on average on the *hard* subset compared to *easy*, even when the required knowledge stays the same, indicating that LLMs are far from robust on geometric knowledge reasoning.

Noisy knowledge does help LLMs solve knowledge crosswords. Despite the existence of irrelevant and confounding information, LLMs do benefit from the prepended noisy knowledge. On average, the W/ KNOWLEDGE settings exhibit a 34.3% FC gain compared to W/O KNOWLEDGE settings across all approaches. This indicates that LLMs possess preliminary abilities to selectively leverage knowledge and information.

Advanced prompting methods show little improvement. Specifically, COT, LTM and COT+SC do not greatly advance performance compared to ZERO-SHOT and FEW-SHOT prompting: the average PC and FC of COT, LTM and COT+SC is 5.6% and 10.4% **less** than those of ZERO-SHOT and FEW-SHOT. This suggests that the left-to-right reasoning patterns employed by these prompting techniques may not be applicable for knowledge crosswords, as these prompting methods fail to induce non-linear reasoning steps for verification and backtracking.

Promoting verification and backtracking improves geometric knowledge reasoning. With GPT-3.5-TURBO, VERIFY-ALL greatly outperforms all baselines with explicit self-verification. Interestingly, after a closer look into the responses, we find that factual errors are rarely detected while the performance gain mainly comes from LLMs proposing more precise answers in a single attempt when specifically asking for fact verification. In addition, while GPT-3.5-TURBO struggles at learning complex reasoning steps required by STAGED PROMPTING, Table 2 shows that STAGED PROMPTING achieves impressive results with GPT-4 and generally outperforms all other methods including VERIFY-ALL in the W/ KNOWLEDGE setting. This indicates that the more elaborate instructions of STAGED PROMPTING work best with more advanced LLMs, as smaller models struggle to grasp these detailed reasoning steps.

6 Analysis

Error Analysis Each knowledge crossword comes with W/ and W/O KNOWLEDGE settings. We conduct error analysis to investigate the impact of noisy passages on LLM problem solving and present in Figure 3. We label each problem based on whether it is answered (in)correctly in W/ KNOWLEDGE (K+(–)) and answered (in)correctly in W/O KNOWLEDGE (N+(–)).

Figure 3 reveals that for easy and medium problems, the main bottleneck is knowledge access since most N– questions are correctly answered under W/ KNOWLEDGE. However, for hard problems, the bottleneck is geometric knowledge reasoning abilities, given that the proportion of N–K– is consistently above 30%, showing that LLMs struggle to reason even when the required knowledge is provided. By including three subsets of varying difficulty in KNOWLEDGE CROSSWORDS, we successfully reveal the multitudes of LLM limitations in knowledge-intensive contexts, disentangling limitations of knowledge and reasoning.

None of the Above To study whether LLMs are subject to "none-of-the-above (NOTA) scenarios", we add an instruction *"Output 'none of the above' if none of the option combinations satisfy all the constraints."* and evaluate the performance of GPT-3.5-TURBO with the FEW-SHOT prompting. Specifically, we experiment with two settings: #1. The correct option is removed from candidates and LLMs should choose to report NOTA; #2. The correct option exists and LLMs should provide answers. Table 3 demonstrates that LLMs struggle

w/ NOTA?	w/ correct?	easy		medium		hard	
		PC	FC	PC	FC	PC	FC
✓	✗	35.7	14.0	35.6	13.5	11.4	3.1
✓	✓	<u>68.1</u>	<u>46.8</u>	<u>69.1</u>	<u>48.3</u>	<u>50.7</u>	<u>20.6</u>
✗	✓	**83.7**	**60.8**	**84.7**	**63.3**	**56.8**	**25.3**

Table 3: FC and PC (%) with FEW-SHOT in the W/O KNOWLEDGE setting. The best results are **bold-faced** and the second-best <u>underlined</u>. "w/ NOTA" denotes where LMs are asked to consider none-of-the-above through instructions and "w/ correct" denotes whether the correct combination is actually provided.

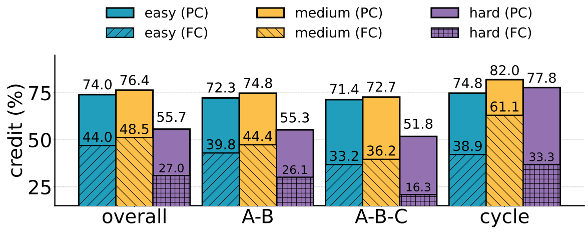

Figure 5: FC and PC (%) for problems with specific structural patterns using CoT. "A-B" denotes two connected blanks, "A-B-C" denotes a chain of three blanks, "cycle" denotes a cycle of three or more blanks, and "overall" denotes the performance on all question graphs.

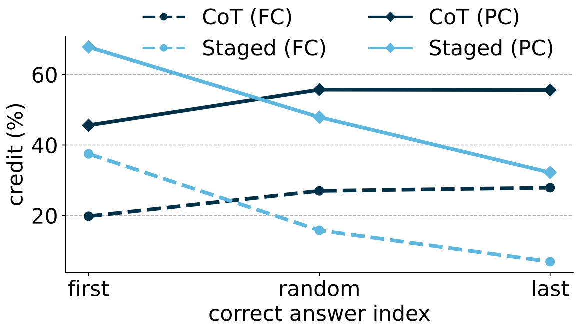

Figure 4: FC and PC (%) under the W/O KNOWLEDGE setting using CoT and STAGED PROMPTING with different orders of options evaluated on the hard subset.

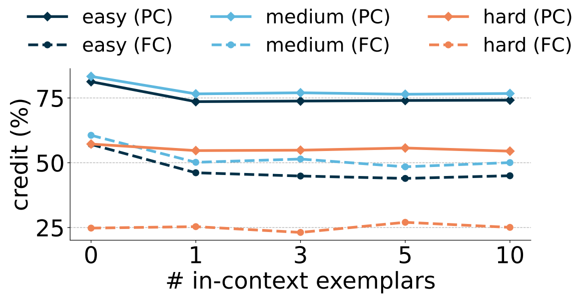

Figure 6: FC and PC (%) using CoT under the W/O KNOWLEDGE setting with an increasing number of in-context exemplars. An increase in the number of exemplars does not necessarily bring performance gain.

with the NOTA scenario, regardless of whether the knowledge crossword is indeed coming without correct answers.

Option Order As our proposed STAGED PROMPTING considers one candidate at one time, we expect that model performance may be worse for problems where the correct answer appears later in the prompt. Figure 4 demonstrates this negative correlation, which could be attributed to LLM hallucination (Ji et al., 2023) and falsely accepting earlier incorrect options. On the other hand, we see an opposite trend in the performance of CoT. This indicates that CoT does not consider the options sequentially and later options might influence the prediction more.

Structural Patterns We investigate whether special structural patterns of blanks in the question graphs might impact LLM performance. We identify three patterns: 1) *A-B*, where two blanks are connected by an edge; 2) *A-B-C*, where three blanks are on a chain; 3) *cycle*, where three more more blanks form a cycle. Figure 5 demonstrates a decrease in performance of *A-B* and *A-B-C* compared to the full dataset, showing a chain of blanks would pose challenges to LLM knowledge reasoning. On the other hand, *cycle* exhibits performance gains: we hypothesize that it has a higher blank-to-constraint ratio (closer to 1:1) than other patterns, which gives LLMs more constraints to work with.

Number of In-Context Exemplars Despite the in-context learning ability demonstrated by LLMs (Brown et al., 2020), we find that more in-context exemplars fail to improve model performance on KNOWLEDGE CROSSWORDS. As presented in Figure 6, for questions with all three difficulty levels, the best performance is achieved at ZERO-SHOT except for the Full-Credit of hard problems. This indicates that left-to-right CoT reasoning could not be adequately learned in context for the problem of knowledge crosswords.

Difficulty of In-Context Exemplars We investigate the correlation between the difficulty of in-context exemplars and model performance by evaluating the performance with 5-shot CoT using 4 different sets of in-context exemplars: *easy*, where all in-context examples come from the easy subset; similarly *medium* and *hard*; *mixed*, where a mixture of 2 easy, 2 medium, and 1 hard examples are

exemplar	test					
	easy		medium		hard	
	PC	FC	PC	FC	PC	FC
easy	73.9	44.4	75.9	47.9	55.2	24.0
medium	**75.4**	**47.9**	**77.2**	**49.7**	55.1	25.9
hard	73.4	43.5	76.2	48.7	55.3	24.0
mixed	74.9	46.0	75.7	47.4	**55.8**	**26.5**

Table 4: FC (%) with CoT using exemplars of varying difficulties under the W/O KNOWLEDGE setting. The best results are in **bold**.

Method	PC	FC
RANDOM	32.8	5.0
ZERO-SHOT	29.6	5.0
FEW-SHOT	35.5	7.0
INSTRUCTION-TUNING	47.3	17.0

Table 5: ZERO-SHOT and FEW-SHOT results are evaluated with LLAMA2-7B on 100 randomly sampled problems without fine-tuning. After INSTRUCTION-TUNING on 1,471 knowledge crosswords, the performance improves.

employed. Table 4 demonstrates that medium or mixed in-context examples work best, while solely employing easy or hard ones is marginally worse. As a result, we follow the *mixed* settings in the main experiments.

Fine-tuning and open-source LMs We additionally evaluate the geometric knowledge reasoning abilities of an open-source language model - LLAMA2-7B with 100 problems randomly selected across all difficulty subsets. Without fine-tuning, LLAMA2-7B demonstrates a performance close to random guess. After instruction-tuning with 1,471 knowledge crosswords randomly selected from all 2,101 questions, the Partial-Credit and Full-Credit become 17.7% and 12.0% higher than ZERO-SHOT prompting as reported in Table 5. This indicates that instruction tuning (Wei et al., 2021) could augment LLMs for solving knowledge crosswords, while to what extent they work with larger LLMs requires further research.

Number of Options As the number of options for each blank increases, the problem becomes harder due to the presence of more confounders. We expect to see a downward trend in model performance when there are more distractors per blank.

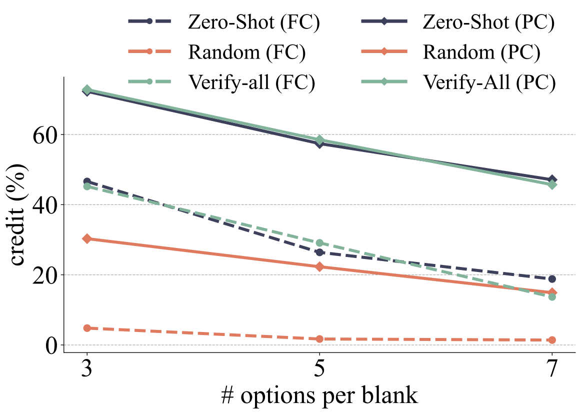

Figure 7: FC and PC (%) evaluated on 292 problems using ZERO-SHOT for increasing number of options per blank. RANDOM denotes the baseline of random guess.

Unsurprisingly, the results in Figure 7 show that the performance is negatively correlated with the number of options per blank. We also observe that performance gap with random guessing is narrowing, suggesting that KNOWLEDGE CROSSWORDS might be difficult for LLMs in the W/O KNOWLEDGE generation setting. (Appendix A)

7 Related Work

Understanding LLM Knowledge LLMs could memorize and encode factual knowledge in model parameters (Petroni et al., 2019; Yu et al., 2023a). As a result, previous research focuses on investigating the extent to which LLMs retrieve and utilize factual knowledge (Yu et al., 2022; Chen et al., 2023a; Mallen et al., 2023). However, the knowledge abilities of LLMs also come with a wide range of limitations such as knowledge update (Hase et al., 2023), irrelevant information (Shi et al., 2023a), and more (Chen et al., 2022; Mruthyunjaya et al., 2023; Kandpal et al., 2023; Sun et al., 2023; Kandpal et al., 2023; Xie et al., 2023; Amayuelas et al., 2023; Wang et al., 2023e; Huang et al., 2023). As a result, various lines of research aim to expand the knowledge abilities of language models, such as better prompting (Press et al., 2023; Sun et al., 2022; Yu et al., 2022; Kojima et al., 2022; Ye and Durrett, 2022), retrieval augmentation (Shi et al., 2023b; Yu et al., 2023b; Asai et al., 2023; Borgeaud et al., 2022; Jiang et al., 2023b), search engine integration (Yu et al., 2022; Press et al., 2023; Kasai et al., 2022; Qin et al., 2023; Vu et al., 2023; Khalifa et al., 2023), and more. While these works primarily focus on evaluating and improving abilities to handle atomic (e.g., open-domain QA) or linear (e.g., multi-hop QA) knowledge, we propose to assess whether LLMs could reason with fact networks bounded by geometric constraints that better align with the structural nature of knowledge.

Reasoning over Knowledge Graphs Simple and complex questions in KBQA have been extensively studied, covering varied tasks including temporal QA (Li et al., 2022; Shang et al., 2022; Chen et al., 2023b; Saxena et al., 2022; Xia et al., 2022; Mei et al., 2022; Ding et al., 2022; Kannen et al., 2022), conversational QA (Ke et al., 2022), general QA (Zhang et al., 2022a; Bai et al., 2022), and more. A myriad of methods have been proposed to tackle these problems, including enhancing reasoning over knowledge graphs (Cao et al., 2022b,a; Ye et al., 2022; Neelam et al., 2022; Xie et al., 2022; Patidar et al., 2023; Zhang et al., 2023; Gupta and Lewis, 2018; Zettlemoyer and Collins, 2009; Cui et al., 2017; Zhong et al., 2017; Shen et al., 2019) and incorporating the generating ability of language models (Liu et al., 2022; Shu et al., 2022; Tang et al., 2022; Hu et al., 2022; Zhang et al., 2022b; Jiang et al., 2023a; Wang et al., 2023b; Kim et al., 2023; Li et al., 2023; Guo et al., 2023; Aglionby and Teufel, 2022; Zhang et al., 2022c). However, it remains underexplored whether LLMs, with the help of its *internal* parametric knowledge, could perform geometric knowledge reasoning with elements similar to these works. We propose KNOWLEDGE CROSSWORDS to investigate LLMs' ability to utilize their linearly acquired knowledge for structured knowledge reasoning scenarios.

Reasoning with Large Language Models LLMs have been evaluated on a myriad of reasoning tasks in an in-context learning setting, including math problems (Ling et al., 2017; Lewkowycz et al., 2022), logical reasoning (Srivastava et al., 2023; Huang et al., 2022), factual knowledge reasoning (Press et al., 2023; Feng et al., 2024), commonsense reasoning (Talmor et al., 2019; Fang et al., 2024; Wang et al., 2023c,d, 2024), and more. Leveraging the in-context learning behavior of LLMs, various prompting techniques (Wei et al., 2022; Zhou et al., 2022; Khot et al., 2022; Wang et al., 2022, 2023a; Schick et al., 2023; Gao et al., 2023) have been proposed to boost the reasoning ability. Specifically, Khot et al. (2022) and Yao et al. (2023) incorporate programs as guides to LLM generation. In this work, we focus on the geometric knowledge reasoning ability of LLMs, which is different from existing left-to-right reasoning patterns, with minimal explicit program-based guidance, involving self-verification, backtracking, and more.

8 Conclusion

We propose KNOWLEDGE CROSSWORDS to investigate LLMs for geometric knowledge reasoning, *i.e.* inferring missing information from an incomplete fact network bounded by geometric constraints. Extensive experiments demonstrate that while existing prompting approaches struggle to solve problems in the *hard* subset, our proposed STAGED PROMPTING and VERIFY-ALL strategies advance model performance while augmenting LLMs with abilities to verify facts, backtrack, and more. Further analysis reveals that LLMs are brittle to "none-of-the-above" scenarios, challenging structural patterns, spurious correlations such as option order, and more.

Limitations

Limited Data Source We construct KNOWLEDGE CROSSWORDS based on only the encyclopedic knowledge graph YAGO, which covers topics on general knowledge about people, cities, countries, movies, and organizations from Wikidata. Since we will make the code publicly available, we leave it to future work on evaluating the geometric reasoning ability of LLMs on different topics with various data sources, such as biomedical knowledge graphs (Chang et al., 2020) and networks in social sciences (Feng et al., 2022).

Answer Uniqueness Due to the incompleteness of knowledge graphs, it is possible that the answer to a problem in KNOWLEDGE CROSSWORDS is not unique. However, such likelihood is presumably low and does not hurt the general evaluation of the geometric reasoning ability of LLMs.

Acknowledgements

This material is based upon work supported by the National Science Foundation under CAREER Grant No. IIS2142739 and NSF Grant No. IIS2203097. We also gratefully acknowledge support from Alfred P. Sloan Foundation Fellowship.

References

Guy Aglionby and Simone Teufel. 2022. Faithful knowledge graph explanations in commonsense question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10811–10817, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*.

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

Yushi Bai, Xin Lv, Juanzi Li, Lei Hou, Yincen Qu, Zelin Dai, and Feiyu Xiong. 2022. Squire: A sequence-to-sequence framework for multi-hop knowledge graph reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1649–1662.

Yuyang Bai, Shangbin Feng, Vidhisha Balachandran, Zhaoxuan Tan, Shiqi Lou, Tianxing He, and Yulia Tsvetkov. 2023. Kgquiz: Evaluating the generalization of encoded knowledge in large language models. *arXiv preprint arXiv:2310.09725*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022a. Kqa pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6101–6119.

Shulin Cao, Jiaxin Shi, Zijun Yao, Xin Lv, Jifan Yu, Lei Hou, Juanzi Li, Zhiyuan Liu, and Jinghui Xiao. 2022b. Program transfer for answering complex questions over knowledge bases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8128–8140.

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Richard Andrew Taylor. 2020. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 167. NIH Public Access.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023a. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. *arXiv preprint arXiv:2310.07289*.

Ziyang Chen, Jinzhi Liao, and Xiang Zhao. 2023b. Multi-granularity temporal question answering over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11378–11392.

Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. Kbqa: Learning question answering over qa corpora and knowledge bases. *Proceedings of the VLDB Endowment*, 10(5).

Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia, and Andrew McCallum. 2022. Knowledge base question answering by case-based reasoning over subgraphs. In *International conference on machine learning*, pages 4777–4793. PMLR.

Wentao Ding, Hao Chen, Huayu Li, and Yuzhong Qu. 2022. Semantic framework based query generation for temporal question answering over knowledge graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1867–1877.

Tianqing Fang, Zeming Chen, Yangqiu Song, and Antoine Bosselut. 2024. Complex reasoning over logical queries on commonsense knowledge graphs. *arXiv preprint arXiv:2403.07398*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Shangbin Feng, Zhaoxuan Tan, Zilong Chen, Ningnan Wang, Peisheng Yu, Qinghua Zheng, Xiaojun Chang, and Minnan Luo. 2022. Par: Political actor representation learning with social context and expert knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12022–12036.

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.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond i.i.d.: Three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, WWW '21, page 3477–3488, New York, NY, USA. Association for Computing Machinery.

Wangzhen Guo, Linyin Luo, Hanjiang Lai, and Jian Yin. 2023. From parse-execute to parse-execute-refine: Improving semantic parser for complex question answering over knowledge base. *arXiv preprint arXiv:2305.03356*.

Nitish Gupta and Mike Lewis. 2018. Neural compositional denotational semantics for question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2152–2161.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2023. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2706–2723.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. 2023. Measuring and manipulating knowledge representations in language models. *arXiv preprint arXiv:2304.00740*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. 2022. Empowering language models with knowledge graph reasoning for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9581.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Qian Huang, Hongyu Ren, and Jure Leskovec. 2022. Few-shot relational reasoning via connection subgraph pretraining. *Advances in Neural Information Processing Systems*, 35:6397–6409.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jinhao Jiang, Kun Zhou, Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2023a. ReasoningLM: Enabling structural subgraph reasoning in pre-trained language models for question answering over knowledge graph. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

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

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

Nithish Kannen, Udit Sharma, Sumit Neelam, Dinesh Khandelwal, Shajith Ikbal, Hima Karanam, and L Venkata Subramaniam. 2022. Targeted extraction of temporal facts from textual resources for improved temporal question answering over knowledge bases. *arXiv preprint arXiv:2203.11054*.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Xirui Ke, Jing Zhang, Xin Lv, Yiqi Xu, Shulin Cao, Cuiping Li, Hong Chen, and Juanzi Li. 2022. Knowledge-augmented self-training of a question rewriter for conversational knowledge base question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1844–1856.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Few-shot reranking for multi-hop QA via language model prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15882–15897, Toronto, Canada. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

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.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. 2023. Few-shot in-context learning on knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6980, Toronto, Canada. Association for Computational Linguistics.

Zixuan Li, Saiping Guan, Xiaolong Jin, Weihua Peng, Yajuan Lyu, Yong Zhu, Long Bai, Wei Li, Jiafeng Guo, and Xueqi Cheng. 2022. Complex evolutional pattern learning for temporal knowledge graph reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 290–296.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Ye Liu, Semih Yavuz, Rui Meng, Dragomir Radev, Caiming Xiong, and Yingbo Zhou. 2022. Uni-parser: Unified semantic parser for question answering on knowledge base and database. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8858–8869.

Linhao Luo, Jiaxin Ju, Bo Xiong, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2023. Chatrule: Mining logical rules with large language models for knowledge graph reasoning. *arXiv preprint arXiv:2309.01538*.

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)*, pages 9802–9822.

Xin Mei, Libin Yang, Xiaoyan Cai, and Zuowei Jiang. 2022. An adaptive logical rule embedding model for inductive reasoning over temporal knowledge graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7304–7316.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Vishwas Mruthyunjaya, Pouya Pezeshkpour, Estevam Hruschka, and Nikita Bhutani. 2023. Rethinking language models as symbolic knowledge graphs. *arXiv preprint arXiv:2308.13676*.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, et al. 2022. Sygma: A system for generalizable and modular question answering over knowledge bases. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3866–3879.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824.

Mayur Patidar, Prayushi Faldu, Avinash Singh, Lovekesh Vig, Indrajit Bhattacharya, and Mausam. 2023. Do I have the knowledge to answer? investigating answerability of knowledge base questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10341–10357, Toronto, Canada. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. " merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *arXiv preprint arXiv:2309.08594*.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. WebCPM: Interactive web search for Chinese long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8968–8988, Toronto, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ray Reagans and Bill McEvily. 2003. Network structure and knowledge transfer: The effects of cohesion and range. *Administrative science quarterly*, 48(2):240–267.

Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-sequence knowledge graph completion and question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2814–2828.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8017–8026.

Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-task learning for conversational question answering over a large-scale knowledge base. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2442–2451.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Harry Shomer, Wei Jin, Wentao Wang, and Jiliang Tang. 2023. Toward degree bias in embedding-based knowledge graph completion. In *Proceedings of the ACM Web Conference 2023*, pages 705–715.

Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. 2022. Tiara: Multi-grained retrieval for robust question answering over large knowledge base. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8108–8121.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Fabian Suchanek, Mehwish Alam, Thomas Bonald, Pierre-Henri Paris, and Jules Soria. 2023. Integrating the wikidata taxonomy into yago. *arXiv preprint arXiv:2308.11884*.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. 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 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Yechun Tang, Xiaoxia Cheng, and Weiming Lu. 2022. Improving complex knowledge base question answering via question-to-action and question-to-question alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 137–147.

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

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Siyuan Wang, Zhongyu Wei, Meng Han, Zhihao Fan, Haijun Shan, Qi Zhang, and Xuanjing Huang. 2023b. Query structure modeling for inductive logical reasoning over knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4706–4718, Toronto, Canada. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023c. Car: Conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13520–13545.

Weiqi Wang, Tianqing Fang, Chunyang Li, Haochen Shi, Wenxuan Ding, Baixuan Xu, Zhaowei Wang, Jiaxin Bai, Xin Liu, Jiayang Cheng, et al. 2024. Candle: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *arXiv preprint arXiv:2401.07286*.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023d. Cat: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13111–13140.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023e. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Jason Wei, Maarten Bosma, Vincent 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. In *International Conference on Learning Representations*.

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.

Yuwei Xia, Mengqi Zhang, Qiang Liu, Shu Wu, and Xiao-Yu Zhang. 2022. Metatkg: Learning evolutionary meta-knowledge for temporal knowledge graph reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7230–7240.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Minghui Xie, Chuzhan Hao, and Peng Zhang. 2022. A sequential flow control framework for multi-hop knowledge base question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8450–8460.

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.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners. *arXiv preprint arXiv:2212.10923*.

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. 2023. Logical reasoning over natural language as knowledge representation: A survey. In *The 61st Annual Meeting Of The 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*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2022. Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6032–6043.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023a. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Pengfei Yu and Heng Ji. 2023. Self information update for large language models through mitigating exposure bias. *arXiv preprint arXiv:2305.18582*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, page 1050–1055. AAAI Press.

Luke S Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 976–984.

Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. 2022a. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5773–5784.

Lingxi Zhang, Jing Zhang, Yanling Wang, Shulin Cao, Xinmei Huang, Cuiping Li, Hong Chen, and Juanzi Li. 2023. FC-KBQA: A fine-to-coarse composition framework for knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1002–1017, Toronto, Canada. Association for Computational Linguistics.

Miao Zhang, Rufeng Dai, Ming Dong, and Tingting He. 2022b. Drlk: dynamic hierarchical reasoning with language model and knowledge graph for question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5123–5133.

X Zhang, A Bosselut, M Yasunaga, H Ren, P Liang, C Manning, and J Leskovec. 2022c. Greaselm: Graph reasoning enhanced language models for question answering. In *International Conference on Representation Learning (ICLR)*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*.

A Discussion

Geometric Reasoning in the W/O KNOWLEDGE generation setting While we mainly focus on solving knowledge crosswords in a multiple-choice setting, it is interesting to evaluate the geometric reasoning ability in the W/O KNOWLEDGE generation setting. Specifically, the problems in KNOWLEDGE CROSSWORDS have unique answers, which should be useful when switching to the W/O KNOWLEDGE generation setting as answer uniqueness makes evaluation easier and makes the problem clearer. Our preliminary experiments show that solving knowledge crosswords in the W/O KNOWLEDGE generation setting is much harder. Considering the model performance in the multiple-choice setting, one method that might be promising is to prompt LLMs themselves to generate candidates for each blank and thereby transform the W/O KNOWLEDGE generation problem into a multiple-choice problem.

Performance gain of VERIFY-ALL While VERIFY-ALL helps LLMs obtain large performance gains in solving knowledge crosswords, it is quite intriguing when investigating where such gains come from. Specifically, in the W/ KNOWLEDGE setting, among all 359 hard problems, we find only 3 problems whose solution with VERIFY-ALL involves detecting errors in verification and repropose candidates. Among the 3 problems, 2 are answered correctly by both VERIFY-ALL and COT, and both methods fail the other problem. This leads to an interesting implication that the performance gain comes from LLMs proposing more precise answers in the first attempt, and that LLMs can jointly consider all constraints rather than consider one by one. We envision the study of such performance gain and the application of the insight as possible future directions.

Application of geometric knowledge reasoning Despite the difficulty of the task, LLMs do show preliminary geometric reasoning ability over incomplete fact network. While such ability still has a long way to achieve perfection, this finding opens up the possibility of using LLMs as flexible relational databases and accessing the parametric knowledge with prompts similar to SQL (structured query language).

Same prompting approach with different LLMs While GPT-3.5-TURBO does not benefit from STAGED PROMPTING, experiments using STAGED PROMPTING with GPT-4 demonstrate impressive results under the W/ KNOWLEDGE setting. Taking a close look at the responses of GPT-3.5-TURBO, we find they fail to follow the reasoning steps presented in the exemplars even if we facilitate the process by guiding the *update* step with program. On the other hand, GPT-4 learn better from exemplars of STAGED PROMPTING with similar settings. This indicates that the success of STAGED PROMPTING relies heavily on the choice of LLMs.

Geometric Knowledge Reasoning vs. Logical Reasoning Logical reasoning, over natural language (Yang et al., 2022, 2023) or logical rules (Pan et al., 2023; Luo et al., 2023), could be confined within pre-defined logic operations set, while geometric knowledge reasoning problems are more flexible and involve diverse logical reasoning types, such as deductive reasoning (applying general knowledge and constraint patterns to deduce the correct answer), abductive reasoning (formulating the most likely answer based on the available clues), and more. Considering the versatile nature of knowledge structure and flexible relational reasoning types involved, geometric knowledge reasoning is reasoning over natural language, without explicit transformation to logic rules as in logical reasoning.

Hyperparameter	Value
degree_c	5, 7, 9
s_G	6, 7, 8, 9, 10, 11
s_B	$[\frac{1}{4} \cdot s_G, \frac{1}{2} \cdot s_G]$
m_r	1.1, 1.2, 1.3
m_b	1, 1.1

Table 6: Hyperparamters for benchmark construction

B KNOWLEDGE CROSSWORDS Details

In this section, we elaborate on the details of benchmark construction and additional experiment details.

B.1 Benchmark Construction Details

1. YAGO filtering
   (a) We remove edges in YAGO with the following relations: (i) Location-related: isLocatedIn, livesIn, happenedIn, diedIn, wasBornIn; (ii) Time-sensitive: worksAt,

playsFor, isAffiliatedTo, isPoliticianOf, isLeaderOf; (iii) Not self-evident: influences, owns, isKnownFor, dealsWith, imports, exports, created, isInterestedIn, dealsWith, isConnectedTo.

   (b) The remaining relations in YAGO are: graduatedFrom, hasMusicalRole, hasAcademicAdvisor, hasChild, wroteMusicFor, hasCapital, actedIn, hasOfficialLanguage, hasWonPrize, hasGender, hasCurrency, directed, isCitizenOf, participatedIn, isMarriedTo, hasNeighbor, edited.

2. Modified k-hop neighborhood
   (a) A center node c is randomly selected from nodes with degree degree_c in filtered YAGO.
   (b) We retrieve a modified 5-hop neighborhood $\mathcal{G}_{\mathcal{N}}^{(c)}$: in each layer, we retain at most 8 nodes. This approach assists us in obtaining a subgraph with a relatively long diameter while avoiding excessive density.
3. Downsample to $\mathcal{G}_{\mathcal{A}}$
   (a) We repeatedly remove nodes from $\mathcal{G}_{\mathcal{N}}^{(c)}$ until the number of nodes in the largest connected component in $\mathcal{G}_{\mathcal{N}}^{(c)}$ is no more than question graph size s_G.
      i. Sort the nodes in $\mathcal{G}_{\mathcal{N}}^{(c)}$ by degree in filtered YAGO in descending order as $\mathbf{v}_{\text{sorted,YAGO}}$.
      ii. Denote *reduce range multiplier* as m_r. Then *reduce range* rr is calculated as $m_r \cdot (|\mathcal{V}_{\mathcal{N}}^{(c)}| - s_G)$ where $\mathcal{V}_{\mathcal{N}}^{(c)}$ is the set of nodes in $\mathcal{G}_{\mathcal{N}}^{(c)}$.
      iii. Randomly pick a node in $\mathbf{v}_{\text{sorted,YAGO}}[(rr-1)/2 : (rr-1)]$ and remove this node from $\mathcal{G}_{\mathcal{N}}^{(c)}$.
   (b) Following the abovementioned approach, we downsample $\mathcal{G}_{\mathcal{N}}^{(c)}$ to $\mathcal{G}_{\mathcal{A}}$ by removing nodes with relatively high degree in filtered YAGO and introduce randomness in this process.
4. Generate blanks to get $\mathcal{G}_{\mathcal{Q}}$
   (a) To mask s_B nodes in $\mathcal{G}_{\mathcal{A}}$ as blanks, denote *blank range multiplier* as m_b and calculate *blank range* br as $s_B \cdot m_b$.
   (b) Sort the nodes in $\mathcal{G}_{\mathcal{N}}^{(c)}$ by degree in $\mathcal{G}_{\mathcal{Q}}$ in descending order as $\mathbf{v}_{\text{sorted},\mathcal{G}_{\mathcal{A}}}$.

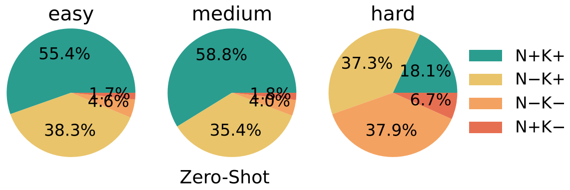

Figure 8: Problem distribution based on the correctness under the W/ KNOWLEDGE and W/O KNOWLEDGE settings using ZERO-SHOT.

   (c) We then randomly select s_B nodes from the first br nodes in $\mathbf{v}_{\text{sorted},\mathcal{G}_{\mathcal{A}}}$ as blanks.

Specifically, the hyperparameters we used for benchmark construction are listed in Tabel 6.

B.2 Relevant Knowledge

In the W/ KNOWLEDGE setting, relevant knowledge is prepended to each problem. Specifically, for each triplet in the constraint, we present four pertinent triplets, with one of them reserved for the triplet containing the correct answer. And the other three are randomly sampled from YAGO following similar method as negative sampling, with priority given to the triplets that satisfy all the criteria (Rule #1, #2, and #3). By sampling relevant triplets in such a way, we provide necessary knowledge (filled with correct answers) as well as confounding knowledge that makes the solving process non-trivial. The sampling of confounding knowledge also simulates the possible information that one may consider when solving the question, with those satisfying all three criteria having the highest likelihood of being considered intuitively.

B.3 Experiment Details

Within 4k-context, in the w/o Relevant Knowledge setting, the numbers of finished responses for easy, medium, and hard questions are 755, 781, and 341 respectively; in the w/ Relevant Knowledge setting, the numbers of finished responses for easy, medium, hard questions are 759, 769 and 328 respectively. The credits are calculated based on these finished responses only.

C Additional Analysis

Error Analysis We provide error analysis conducted with ZERO-SHOT here for reference. Results in Figure 8 shows that all three methods share similar trends.

Number of Blanks We study the impact of the number of blanks in the problem on the model performance. Specifically, we randomly select 100 problems with three blanks (# of blanks = 3) from all three difficulty levels and construct two additional versions of these problems by filling in one (# of blanks = 2) or two (# of blanks = 1) answers to the blanks. We evaluate the performance of various methods (ZERO-SHOT, FEW-SHOT, COT, VERIFY-ALL) and two settings (W/ KNOWLEDGE or W/O KNOWLEDGE) on these three versions of the dataset and present the results in Table 7. We find that ChatGPT performs well on the simpler (# of blanks = 1) version, demonstrating a strong knowledge ability. However, its performance suffers when the required reasoning steps increase and the geometric structures involved become more complex.

D Qualitative analysis

In this section, we provide examples of knowledge crosswords that GPT-3.5-TURBO answers correctly or wrongly using STAGED PROMPTING and VERIFY-ALL. In-context exemplars are omitted in this section to save space and can be found in Appendix E. Table 8 and Table 9 show results using STAGED PROMPTING; Table 10, Table 11 and Table 12 show results using VERIFY-ALL.

E Prompts

We list the prompts for all experiments of Tables 2 and 5 in Tables 13, 14, 15, 16, 17, 18, 19.

# of blanks	Methods	W/ KNOWLEDGE						W/O KNOWLEDGE					
		easy		medium		hard		easy		medium		hard	
		PC	FC	PC	FC	PC	FC	PC	FC	PC	FC	PC	FC
3	ZERO-SHOT	98.0	95.0	98.3	95.0	77.7	51.0	85.0	64.0	87.3	68.0	59.3	21.0
3	FEW-SHOT	98.7	96.0	98.0	95.0	77.3	44.0	83.0	57.0	84.0	61.0	52.0	12.0
3	COT	93.7	87.0	94.3	90.0	78.0	52.0	73.7	45.0	79.3	49.0	49.3	18.0
3	VERIFY-ALL	99.3	98.0	98.0	94.0	81.0	54.0	86.3	64.0	88.7	68.0	57.0	21.0
2	ZERO-SHOT	96.0	94.0	93.5	90.0	78.0	67.0	84.5	73.0	82.5	68.0	61.0	34.0
2	FEW-SHOT	97.5	95.0	96.5	93.0	85.5	74.0	82.5	68.0	86.0	74.0	66.0	44.0
2	COT	91.5	85.0	93.0	89.0	82.0	71.0	78.5	62.0	78.5	64.0	58.5	41.0
2	VERIFY-ALL	98.5	97.0	98.0	96.0	89.0	81.0	87.0	79.0	87.0	76.0	60.5	45.0
1	ZERO-SHOT	96.0	96.0	91.0	91.0	76.0	76.0	92.0	92.0	88.0	88.0	63.0	63.0
1	FEW-SHOT	99.0	99.0	97.0	97.0	87.0	87.0	92.0	92.0	92.0	92.0	64.0	64.0
1	COT	97.0	97.0	93.0	93.0	83.0	83.0	92.0	92.0	90.0	90.0	59.0	59.0
1	VERIFY-ALL	100.0	100.0	97.0	97.0	91.0	91.0	92.0	92.0	87.0	87.0	69.0	69.0

Table 7: FC and PC with GPT-3.5-TURBO on 100 randomly sampled problems with three blanks. We fill in (3 - # of blanks) blanks in each problem and the LM is tasked with figuring out the remaining blanks.

Table 8: Response using STAGED PROMPTING where the answers are correct.

Prompt	Response
[STAGED PROMPTING exemplars] Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (June Allyson, is married to, Dick Powell); (Bradley Cooper, is married to, Jennifer Esposito); (Charlie Chaplin, is married to, Oona O'Neill); (Joan Blondell, is married to, Dick Powell); (Joan Blondell, acted in, Support Your Local Gunfighter); (Suzanne Pleshette, acted in, Support Your Local Gunfighter); (James Garner, acted in, Support Your Local Gunfighter); (Jack Elam, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, June Allyson); (Bradley Cooper, is married to, Jennifer Esposito); (Charlie Chaplin, is married to, Oona O'Neill); (Dick Powell, is married to, Joan Blondell); (Dick Powell, has gender, male); (Eric Maskin, has gender, male); (Charlie Chaplin, has gender, male); (Robert Bly, has gender, male); (Eric Maskin, has gender, male); (Robert Bly, has gender, male); (Charlie Chaplin, has gender, male); (Borislav Mikhailov, has gender, male); (Charlie Chaplin, has gender, male); (Eric Maskin, has gender, male); (Robert Bly, has gender, male); (Cole Tinkler, has gender, male). Constraints: (blank 2, is married to, Dick Powell); (blank 2, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, blank 2); (Dick Powell, has gender, blank 1); (Borislav Mikhailov, has gender, blank 1); (Cole Tinkler, has gender, blank 1). Options: blank 1: A. female, B. male blank 2: A. Suzanne Pleshette, B. Joan Blondell, C. James Garner Answer:	Stage 1 - status: Current answer: None. Remaining constraints containing blanks: (blank 2, is married to, Dick Powell); (blank 2, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, blank 2); (Dick Powell, has gender, blank 1); (Borislav Mikhailov, has gender, blank 1); (Cole Tinkler, has gender, blank 1). Stage 1 - solve: From (Dick Powell, has gender, blank 1), candidate for blank 1: B. male. Stage 1 - status update: Filled remaining constraints with current answer: (Dick Powell, has gender, male); (Borislav Mikhailov, has gender, male); (Cole Tinkler, has gender, male). Updated remaining constraints containing blanks: (blank 2, is married to, Dick Powell); (blank 2, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, blank 2). Stage1 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 2 - status: Current answer: blank 1: B. male. Remaining constraints containing blanks: (blank 2, is married to, Dick Powell); (blank 2, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, blank 2). Stage 2 - solve: From (blank 2, is married to, Dick Powell), candidate for blank 2: B. Joan Blondell. Stage 2 - status update: Filled remaining constraints with current answer: (Joan Blondell, is married to, Dick Powell); (Joan Blondell, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, Joan Blondell). Updated remaining constraints containing blanks: Stage2 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 3 - status: Current answer: blank 1: B. male, blank 2: B. Joan Blondell. Remaining constraints with blanks: None. Stage 3 - solve: No more remaining constraints with blank. Final answer: blank 1: B, blank 2: B ✓

Table 9: Response using STAGED PROMPTING where the answers are wrong.

Prompt	**Response**
[STAGED PROMPTING exemplars] Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Paz Vega, acted in, The Human Contract); (Joe Roberts, acted in, Our Hospitality); (Paz Vega, acted in, The Six Wives of Henry Lefay); (Paz Vega, acted in, The Spirit (film)); (Charlie Chaplin, directed, A Woman of Paris); (Charlie Chaplin, directed, The Champion (1915 film)); (Charlie Chaplin, directed, Limelight (1952 film)); (Jada Pinkett Smith, directed, The Human Contract); (Johnathon Schaech, acted in, Prom Night (2008 film)); (Joe Roberts, acted in, Our Hospitality); (Brittany Snow, acted in, Prom Night (2008 film)); (Idris Elba, acted in, Prom Night (2008 film)); (Joe Roberts, acted in, The Haunted House (1921 film)); (Idris Elba, acted in, The Human Contract); (Joe Roberts, acted in, Our Hospitality); (William Hopper, acted in, The Bad Seed (1956 film)). Constraints: (Paz Vega, acted in, blank 1); (Jada Pinkett Smith, directed, blank 1); (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, blank 1). Options: blank 1: A. The Human Contract, B. The Spirit (film), C. The Six Wives of Henry Lefay blank 2: A. Johnathon Schaech, B. Idris Elba, C. Brittany Snow Answer:	Stage 1 - status: Current answer: None. Remaining constraints containing blanks: (Paz Vega, acted in, blank 1); (Jada Pinkett Smith, directed, blank 1); (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, blank 1). Stage 1 - solve: From (Paz Vega, acted in, blank 1), candidate for blank 1: A. The Human Contract. Stage 1 - status update: Filled remaining constraints with current answer: (Paz Vega, acted in, The Human Contract); (Jada Pinkett Smith, directed, The Human Contract). Updated remaining constraints containing blanks: (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, The Human Contract). Stage1 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 2 - status: Current answer: blank 1: A. The Human Contract. Remaining constraints containing blanks: (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, The Human Contract). Stage 2 - solve: From (blank 2, acted in, Prom Night (2008 film)), candidate for blank 2: A. Johnathon Schaech. Stage 2 - status update: Filled remaining constraints with current answer: (Johnathon Schaech, acted in, Prom Night (2008 film)); (Johnathon Schaech, acted in, The Human Contract). Updated remaining constraints containing blanks: Stage2 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 3 - status: Current answer: blank 1: A. The Human Contract, blank 2: A. Johnathon Schaech. Remaining constraints with blanks: None. Stage 3 - solve: No more remaining constraints with blank. Final answer: blank 1: A, blank 2: A ✗(Correct answer: blank 1: A, blank 2: B)

Table 10: Response using VERIFY-ALL where the answers are correct and involves error detection and backtracking.

Prompt	**Response**
[VERIFY-ALL exemplars] Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (June Allyson, is married to, Dick Powell); (Bradley Cooper, is married to, Jennifer Esposito); (Charlie Chaplin, is married to, Oona O'Neill); (Joan Blondell, is married to, Dick Powell); (Joan Blondell, acted in, Support Your Local Gunfighter); (Suzanne Pleshette, acted in, Support Your Local Gunfighter); (James Garner, acted in, Support Your Local Gunfighter); (Jack Elam, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, June Allyson); (Bradley Cooper, is married to, Jennifer Esposito); (Charlie Chaplin, is married to, Oona O'Neill); (Dick Powell, is married to, Joan Blondell); (Dick Powell, has gender, male); (Eric Maskin, has gender, male); (Charlie Chaplin, has gender, male); (Robert Bly, has gender, male); (Eric Maskin, has gender, male); (Robert Bly, has gender, male); (Charlie Chaplin, has gender, male); (Borislav Mikhailov, has gender, male); (Charlie Chaplin, has gender, male); (Eric Maskin, has gender, male); (Robert Bly, has gender, male); (Cole Tinkler, has gender, male). Constraints: (blank 2, is married to, Dick Powell); (blank 2, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, blank 2); (Dick Powell, has gender, blank 1); (Borislav Mikhailov, has gender, blank 1); (Cole Tinkler, has gender, blank 1). Options: blank 1: A. female, B. male blank 2: A. Suzanne Pleshette, B. Joan Blondell, C. James Garner Answer:	Candidate answer: blank 1: A. female, blank 2: B. Joan Blondell; Filled constraints with candidate answer: (Joan Blondell, is married to, Dick Powell); (Joan Blondell, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, Joan Blondell); (Dick Powell, has gender, female); (Borislav Mikhailov, has gender, female); (Cole Tinkler, has gender, female); Verification: Does error occur in filled constraints with candidate answer? (Dick Powell, has gender, female) is incorrect. Candidate answer: blank 1: B. male, blank 2: B. Joan Blondell; Filled constraints with candidate answer: (Joan Blondell, is married to, Dick Powell); (Joan Blondell, acted in, Support Your Local Gunfighter); (Dick Powell, is married to, Joan Blondell); (Dick Powell, has gender, male); (Borislav Mikhailov, has gender, male); (Cole Tinkler, has gender, male); Verification: Does error occur in filled constraints with candidate answer? No. Therefore, blank 1: B, blank 2: B. ✓

Table 11: Response using VERIFY-ALL where the answers are correct correctly in a single trial.

Prompt	**Response**
[VERIFY-ALL exemplars] Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Paz Vega, acted in, The Human Contract); (Joe Roberts, acted in, Our Hospitality); (Paz Vega, acted in, The Six Wives of Henry Lefay); (Paz Vega, acted in, The Spirit (film)); (Charlie Chaplin, directed, A Woman of Paris); (Charlie Chaplin, directed, The Champion (1915 film)); (Charlie Chaplin, directed, Limelight (1952 film)); (Jada Pinkett Smith, directed, The Human Contract); (Johnathon Schaech, acted in, Prom Night (2008 film)); (Joe Roberts, acted in, Our Hospitality); (Brittany Snow, acted in, Prom Night (2008 film)); (Idris Elba, acted in, Prom Night (2008 film)); (Joe Roberts, acted in, The Haunted House (1921 film)); (Idris Elba, acted in, The Human Contract); (Joe Roberts, acted in, Our Hospitality); (William Hopper, acted in, The Bad Seed (1956 film)). Constraints: (Paz Vega, acted in, blank 1); (Jada Pinkett Smith, directed, blank 1); (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, blank 1). Options: blank 1: A. The Human Contract, B. The Spirit (film), C. The Six Wives of Henry Lefay blank 2: A. Johnathon Schaech, B. Idris Elba, C. Brittany Snow Answer:	Candidate answer: blank 1: A. The Human Contract, blank 2: B. Idris Elba; Filled constraints with candidate answer: (Paz Vega, acted in, The Human Contract); (Jada Pinkett Smith, directed, The Human Contract); (Idris Elba, acted in, Prom Night (2008 film)); (Idris Elba, acted in, The Human Contract); Verification: Does error occur in filled constraints with candidate answer? No. Therefore, blank 1: A, blank 2: B. ✓

Table 12: Response using VERIFY-ALL where the answers are wrong.

Prompt	**Response**
[VERIFY-ALL exemplars] Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Colleen Camp, has gender, female); (Robert Bly, has gender, male); (Eric Maskin, has gender, male); (Charlie Chaplin, has gender, male); (Grace Zabriskie, has gender, female); (Robert Bly, has gender, male); (Eric Maskin, has gender, male); (Charlie Chaplin, has gender, male); (Tim Conway, acted in, The Private Eyes (1980 film)); (Don Knotts, acted in, The Private Eyes (1980 film)); (Bernard Fox (actor), acted in, The Private Eyes (1980 film)); (Grace Zabriskie, acted in, The Private Eyes (1980 film)); (Charlie Chaplin, has gender, male); (Eric Maskin, has gender, male); (Diane Baker, has gender, female); (Robert Bly, has gender, male); (Jane Seymour, has gender, female); (Eric Maskin, has gender, male); (Robert Bly, has gender, male); (Charlie Chaplin, has gender, male). Constraints: (Colleen Camp, has gender, blank 2); (blank 1, has gender, blank 2); (blank 1, acted in, The Private Eyes (1980 film)); (Diane Baker, has gender, blank 2); (Jane Seymour, has gender, blank 2). Options: blank 1: A. Bernard Fox (actor), B. Grace Zabriskie, C. Don Knotts blank 2: A. female, B. male Answer:	Candidate answer: blank 1: C. Don Knotts, blank 2: A. female; Filled constraints with candidate answer: (Colleen Camp, has gender, female); (Don Knotts, has gender, female); (Don Knotts, acted in, The Private Eyes (1980 film)); (Diane Baker, has gender, female); (Jane Seymour, has gender, female); Verification: Does error occur in filled constraints with candidate answer? No. Therefore, blank 1: C, blank 2: A. ✗ (Correct answer: blank 1: B, blank 2: A)

Table 13: Prompts used in main experiments with exemplars and an example question. “Knowledge” for each problem is only applicable in the “W/ KNOWLEDGE” setting.

An example of knowledge crossword
Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Desired format: blank i: Z ... Constraints: (Paz Vega, acted in, blank 1); (Jada Pinkett Smith, directed, blank 1); (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, blank 1). Options: blank 1: A. The Human Contract, B. The Spirit (film), C. The Six Wives of Henry Lefay blank 2: A. Johnathon Schaech, B. Idris Elba, C. Brittany Snow Answer: blank 1: A, blank 2: B

Prompt	**Response**
UPPERBOUND	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Desired format: blank i: Z ... Knowledge: (Paz Vega, acted in, The Human Contract); (Jada Pinkett Smith, directed, The Human Contract); (Idris Elba, acted in, Prom Night (2008 film)); (Idris Elba, acted in, The Human Contract). Constraints: (Paz Vega, acted in, blank 1); (Jada Pinkett Smith, directed, blank 1); (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, blank 1). Options: blank 1: A. The Human Contract, B. The Spirit (film), C. The Six Wives of Henry Lefay blank 2: A. Johnathon Schaech, B. Idris Elba, C. Brittany Snow Answer:
ZERO-SHOT	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Paz Vega, acted in, The Human Contract); (Joe Roberts, acted in, Our Hospitality); (Paz Vega, acted in, The Six Wives of Henry Lefay); (Paz Vega, acted in, The Spirit (film)); (Charlie Chaplin, directed, A Woman of Paris); (Charlie Chaplin, directed, The Champion (1915 film)); (Charlie Chaplin, directed, Limelight (1952 film)); (Jada Pinkett Smith, directed, The Human Contract); (Johnathon Schaech, acted in, Prom Night (2008 film)); (Joe Roberts, acted in, Our Hospitality); (Brittany Snow, acted in, Prom Night (2008 film)); (Idris Elba, acted in, Prom Night (2008 film)); (Joe Roberts, acted in, The Haunted House (1921 film)); (Idris Elba, acted in, The Human Contract); (Joe Roberts, acted in, Our Hospitality); (William Hopper, acted in, The Bad Seed (1956 film)). (Optional) Desired format: blank i: Z Constraints: (Paz Vega, acted in, blank 1); (Jada Pinkett Smith, directed, blank 1); (blank 2, acted in, Prom Night (2008 film)); (blank 2, acted in, blank 1). Options: blank 1: A. The Human Contract, B. The Spirit (film), C. The Six Wives of Henry Lefay blank 2: A. Johnathon Schaech, B. Idris Elba, C. Brittany Snow Answer:

Table 14: Prompts used in main experiments with exemplars and an example question. "Knowledge" for each problem is only applicable in the "W/ KNOWLEDGE" setting.

FEW-SHOT	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Charlton Heston, acted in, True Lies); (Eliza Dushku, acted in, True Lies); (Tom Arnold (actor), acted in, True Lies); (Bill Paxton, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)); (Stephen Collins, acted in, Chiefs (miniseries)); (Paul Sorvino, acted in, Chiefs (miniseries)); (Danny Glover, acted in, Chiefs (miniseries)). (Optional) Constraints: (blank 1, acted in, True Lies); (blank 1, acted in, Chiefs (miniseries)). Options: blank 1: A. Bill Paxton, B. Charlton Heston, C. Paul Sorvino Answer: blank 1: B Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Neighbors (1920 film)); (Taye Diggs, acted in, Rent (film)); (Joe Roberts, acted in, Three Ages); (Bradley Cooper, is married to, Jennifer Esposito); (Taye Diggs, is married to, Idina Menzel); (Charlie Chaplin, is married to, Mildred Harris); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Idina Menzel, acted in, Enchanted (film)); (Idina Menzel, acted in, Rent (film)); (Joe Roberts, acted in, Neighbors (1920 film)); (Joe Roberts, acted in, Three Ages); (Idina Menzel, is married to, Taye Diggs); (Bradley Cooper, is married to, Jennifer Esposito); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Charlie Chaplin, is married to, Mildred Harris). (Optional) Constraints: (blank 1, acted in, blank 2); (blank 1, is married to, Idina Menzel); (Idina Menzel, acted in, blank 2); (Idina Menzel, is married to, blank 1). Options: blank 1: A. Kelly LeBrock, B. Napoleon, C. Taye Diggs blank 2: A. Halloweentown High, B. Magnolia (film), C. Rent (film) Answer: blank 1: C, blank 2: C Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Andy García, acted in, Smokin' Aces); (Andy García, acted in, Ocean's Thirteen); (Andy García, acted in, The Untouchables (film)); (Andy García, acted in, The Pink Panther 2); (Jeremy Piven, acted in, Smokin' Aces); (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Neighbors (1920 film)); (Virginia Madsen, acted in, Scooby-Doo! in Where's My Mummy?); (Jeremy Piven, acted in, Scooby-Doo! in Where's My Mummy?); (Mindy Cohn, acted in, Scooby-Doo! in Where's My Mummy?); (Grey DeLisle, acted in, Scooby-Doo! in Where's My Mummy?). (Optional) Constraints: (Andy García, acted in, blank 1); (blank 2, acted in, blank 1); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Options: blank 1: A. Things to Do in Denver When You're Dead, B. Smokin' Aces, C. Beverly Hills Chihuahua blank 2: A. Ron Perlman, B. Casey Kasem, C. Jeremy Piven Answer: blank 1: B, blank 2: C Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Miriam Margolyes, acted in, Harry Potter (film series)); (David Thewlis, acted in, Harry Potter (film series)); (John Cleese, acted in, Harry Potter (film series)); (Richard Harris, acted in, Harry Potter (film series)); (Joe Roberts, acted in, Three Ages); (Maureen Lipman, acted in, A Little Princess (1986 TV serial)); (Nigel Havers, acted in, A Little Princess (1986 TV serial)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)). (Optional) Constraints: (blank 1, acted in, Harry Potter (film series)); (blank 1, acted in, A Little Princess (1986 TV serial)). Options: blank 1: A. Miriam Margolyes, B. Maggie Smith, C. Emma Watson Answer: blank 1: A Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Joe Roberts, acted in, Neighbors (1920 film)); (Dinah Sheridan, acted in, The Railway Children (film)); (Joe Roberts, acted in, Three Ages); (Dinah Sheridan, acted in, 29 Acacia Avenue); (Dinah Sheridan, is married to, John Merivale); (Charlie Chaplin, is married to, Mildred Harris); (Dinah Sheridan, is married to, Jimmy Hanley); (Bradley Cooper, is married to, Jennifer Esposito); (Joe Roberts, acted in, Neighbors (1920 film)); (Jimmy Hanley, acted in, 29 Acacia Avenue); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Our Hospitality); (Charlie Chaplin, is married to, Mildred Harris); (Dinah Sheridan, is married to, Jimmy Hanley); (Charlie Chaplin, is married to, Oona O'Neill); (Dinah Sheridan, is married to, John Merivale); (Charlie Chaplin, is married to, Mildred Harris); (Charlie Chaplin, is married to, Oona O'Neill);(John Merivale, is married to, Jan Sterling); (Paul Douglas (actor), is married to, Jan Sterling). (Optional) Constraints: (Dinah Sheridan, acted in, blank 2); (Dinah Sheridan, is married to, blank 1); (blank 1, acted in, blank 2); (Dinah Sheridan, is married to, blank 3); (blank 3, is married to, Jan Sterling). Options: blank 1: A. Liam Neeson, B. Jimmy Hanley, C. Nancy Wilson (rock musician) blank 2: A. Courage of Lassie, B. 29 Acacia Avenue, C. Listen to Me (film) blank 3: A. José María Aznar, B. John Merivale, C. Prince Harald of Denmark Answer: blank 1: B, blank 2: B, blank 3: B [ZERO-SHOT prompt]

Table 15: Prompts used in main experiments with exemplars and an example question. “Knowledge” for each problem is only applicable in the “W/ KNOWLEDGE” setting.

COT	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Charlton Heston, acted in, True Lies); (Eliza Dushku, acted in, True Lies); (Tom Arnold (actor), acted in, True Lies); (Bill Paxton, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)); (Stephen Collins, acted in, Chiefs (miniseries)); (Paul Sorvino, acted in, Chiefs (miniseries)); (Danny Glover, acted in, Chiefs (miniseries)). (Optional) Constraints: (blank 1, acted in, True Lies); (blank 1, acted in, Chiefs (miniseries)). Options: blank 1: A. Bill Paxton, B. Charlton Heston, C. Paul Sorvino Answer: (Charlton Heston, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)). Therefore, blank 1: B Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Neighbors (1920 film)); (Taye Diggs, acted in, Rent (film)); (Joe Roberts, acted in, Three Ages); (Bradley Cooper, is married to, Jennifer Esposito); (Taye Diggs, is married to, Idina Menzel); (Charlie Chaplin, is married to, Mildred Harris); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Idina Menzel, acted in, Enchanted (film)); (Idina Menzel, acted in, Rent (film)); (Joe Roberts, acted in, Neighbors (1920 film)); (Joe Roberts, acted in, Three Ages); (Idina Menzel, is married to, Taye Diggs); (Bradley Cooper, is married to, Jennifer Esposito); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Charlie Chaplin, is married to, Mildred Harris). (Optional) Constraints: (blank 1, acted in, blank 2); (blank 1, is married to, Idina Menzel); (Idina Menzel, acted in, blank 2); (Idina Menzel, is married to, blank 1). Options: blank 1: A. Kelly LeBrock, B. Napoleon, C. Taye Diggs blank 2: A. Halloweentown High, B. Magnolia (film), C. Rent (film) Answer: (Taye Diggs, acted in, Rent (film)); (Taye Diggs, is married to, Idina Menzel); (Idina Menzel, acted in, Rent (film)). Therefore, blank 1: C, blank 2: C Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Andy García, acted in, Smokin’ Aces); (Andy García, acted in, Ocean’s Thirteen); (Andy García, acted in, The Untouchables (film)); (Andy García, acted in, The Pink Panther 2); (Jeremy Piven, acted in, Smokin’ Aces); (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Neighbors (1920 film)); (Virginia Madsen, acted in, Scooby-Doo! in Where’s My Mummy?); (Jeremy Piven, acted in, Scooby-Doo! in Where’s My Mummy?); (Mindy Cohn, acted in, Scooby-Doo! in Where’s My Mummy?); (Grey DeLisle, acted in, Scooby-Doo! in Where’s My Mummy?). (Optional) Constraints: (Andy García, acted in, blank 1); (blank 2, acted in, blank 1); (blank 2, acted in, Scooby-Doo! in Where’s My Mummy?). Options: blank 1: A. Things to Do in Denver When You’re Dead, B. Smokin’ Aces, C. Beverly Hills Chihuahua blank 2: A. Ron Perlman, B. Casey Kasem, C. Jeremy Piven Answer: (Andy García, acted in, Smokin’ Aces); (Jeremy Piven, acted in, Smokin’ Aces); (Jeremy Piven, acted in, Scooby-Doo! in Where’s My Mummy?). Therefore, blank 1: B, blank 2: C Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Miriam Margolyes, acted in, Harry Potter (film series)); (David Thewlis, acted in, Harry Potter (film series)); (John Cleese, acted in, Harry Potter (film series)); (Richard Harris, acted in, Harry Potter (film series)); (Joe Roberts, acted in, Three Ages); (Maureen Lipman, acted in, A Little Princess (1986 TV serial)); (Nigel Havers, acted in, A Little Princess (1986 TV serial)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)). (Optional) Constraints: (blank 1, acted in, Harry Potter (film series)); (blank 1, acted in, A Little Princess (1986 TV serial)). Options: blank 1: A. Miriam Margolyes, B. Maggie Smith, C. Emma Watson Answer: (Miriam Margolyes, acted in, Harry Potter (film series)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)). Therefore, blank 1: A Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Joe Roberts, acted in, Neighbors (1920 film)); (Dinah Sheridan, acted in, The Railway Children (film)); (Joe Roberts, acted in, Three Ages); (Dinah Sheridan, acted in, 29 Acacia Avenue); (Dinah Sheridan, is married to, John Merivale); (Charlie Chaplin, is married to, Mildred Harris); (Dinah Sheridan, is married to, Jimmy Hanley); (Bradley Cooper, is married to, Jennifer Esposito); (Joe Roberts, acted in, Neighbors (1920 film)); (Jimmy Hanley, acted in, 29 Acacia Avenue); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Our Hospitality); (Charlie Chaplin, is married to, Mildred Harris); (Dinah Sheridan, is married to, Jimmy Hanley); (Charlie Chaplin, is married to, Oona O’Neill); (Dinah Sheridan, is married to, John Merivale); (Charlie Chaplin, is married to, Mildred Harris); (Charlie Chaplin, is married to, Oona O’Neill);(John Merivale, is married to, Jan Sterling); (Paul Douglas (actor), is married to, Jan Sterling). (Optional) Constraints: (Dinah Sheridan, acted in, blank 2); (Dinah Sheridan, is married to, blank 1); (blank 1, acted in, blank 2); (Dinah Sheridan, is married to, blank 3); (blank 3, is married to, Jan Sterling). Options: blank 1: A. Liam Neeson, B. Jimmy Hanley, C. Nancy Wilson (rock musician) blank 2: A. Courage of Lassie, B. 29 Acacia Avenue, C. Listen to Me (film) blank 3: A. José María Aznar, B. John Merivale, C. Prince Harald of Denmark Answer: (Dinah Sheridan, acted in, 29 Acacia Avenue); (Dinah Sheridan, is married to, Jimmy Hanley); (Jimmy Hanley, acted in, 29 Acacia Avenue); (Dinah Sheridan, is married to, John Merivale); (John Merivale, is married to, Jan Sterling). Therefore, blank 1: B, blank 2: B, blank 3: B [ZERO-SHOT prompt]

Table 16: Prompts used in main experiments with exemplars and an example question. "Knowledge" for each problem is only applicable in the "W/ KNOWLEDGE" setting.

LTM	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Charlton Heston, acted in, True Lies); (Eliza Dushku, acted in, True Lies); (Tom Arnold (actor), acted in, True Lies); (Bill Paxton, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)); (Stephen Collins, acted in, Chiefs (miniseries)); (Paul Sorvino, acted in, Chiefs (miniseries)); (Danny Glover, acted in, Chiefs (miniseries)). (Optional) Constraints: (blank 1, acted in, True Lies); (blank 1, acted in, Chiefs (miniseries)). Options: blank 1: A. Bill Paxton, B. Charlton Heston, C. Paul Sorvino Answer: Considering (blank 1, acted in, True Lies), maybe blank 1: A, or blank 1: B; considering (blank 1, acted in, True Lies), (blank 1, acted in, Chiefs (miniseries)), maybe blank 1: B. Therefore, blank 1: B Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Neighbors (1920 film)); (Taye Diggs, acted in, Rent (film)); (Joe Roberts, acted in, Three Ages); (Bradley Cooper, is married to, Jennifer Esposito); (Taye Diggs, is married to, Idina Menzel); (Charlie Chaplin, is married to, Mildred Harris); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Idina Menzel, acted in, Enchanted (film)); (Idina Menzel, acted in, Rent (film)); (Joe Roberts, acted in, Neighbors (1920 film)); (Joe Roberts, acted in, Three Ages); (Idina Menzel, is married to, Taye Diggs); (Bradley Cooper, is married to, Jennifer Esposito); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Charlie Chaplin, is married to, Mildred Harris). (Optional) Constraints: (blank 1, acted in, blank 2); (blank 1, is married to, Idina Menzel); (Idina Menzel, acted in, blank 2); (Idina Menzel, is married to, blank 1). Options: blank 1: A. Kelly LeBrock, B. Napoleon, C. Taye Diggs blank 2: A. Halloweentown High, B. Magnolia (film), C. Rent (film) Answer: Considering (blank 1, acted in, blank 2), maybe blank 1: C and blank 2: C; considering (blank 1, acted in, blank 2), (blank 1, is married to, Idina Menzel), maybe blank 1: C and blank 2: C; considering (blank 1, acted in, blank 2), (blank 1, is married to, Idina Menzel), (Idina Menzel, acted in, blank 2), maybe blank 1: C and blank 2: C; considering (blank 1, acted in, blank 2), (blank 1, is married to, Idina Menzel), (Idina Menzel, acted in, blank 2), (Idina Menzel, is married to, blank 1), maybe blank 1: C and blank 2: C; Therefore, blank 1: C, blank 2: C Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Andy García, acted in, Smokin' Aces); (Andy García, acted in, Ocean's Thirteen); (Andy García, acted in, The Untouchables (film)); (Andy García, acted in, The Pink Panther 2); (Jeremy Piven, acted in, Smokin' Aces); (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Neighbors (1920 film)); (Virginia Madsen, acted in, Scooby-Doo! in Where's My Mummy?); (Jeremy Piven, acted in, Scooby-Doo! in Where's My Mummy?); (Mindy Cohn, acted in, Scooby-Doo! in Where's My Mummy?); (Grey DeLisle, acted in, Scooby-Doo! in Where's My Mummy?). (Optional) Constraints: (Andy García, acted in, blank 1); (blank 2, acted in, blank 1); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Options: blank 1: A. Things to Do in Denver When You're Dead, B. Smokin' Aces, C. Beverly Hills Chihuahua blank 2: A. Ron Perlman, B. Casey Kasem, C. Jeremy Piven Answer: Considering (Andy García, acted in, blank 1), maybe blank 1: A, or blank 1: B, or blank 1: C; considering (Andy García, acted in, blank 1), (blank 2, acted in, blank 1), maybe blank 1: B and blank 2: C; considering (Andy García, acted in, blank 1), (blank 2, acted in, blank 1), (blank 2, acted in, Scooby-Doo! in Where's My Mummy?), maybe blank 1: B and blank 2: C. Therefore, blank 1: B, blank 2: C Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Miriam Margolyes, acted in, Harry Potter (film series)); (David Thewlis, acted in, Harry Potter (film series)); (John Cleese, acted in, Harry Potter (film series)); (Richard Harris, acted in, Harry Potter (film series)); (Joe Roberts, acted in, Three Ages); (Maureen Lipman, acted in, A Little Princess (1986 TV serial)); (Nigel Havers, acted in, A Little Princess (1986 TV serial)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)). (Optional) Constraints: (blank 1, acted in, Harry Potter (film series)); (blank 1, acted in, A Little Princess (1986 TV serial)). Options: blank 1: A. Miriam Margolyes, B. Maggie Smith, C. Emma Watson Answer: Considering (blank 1, acted in, Harry Potter (film series)), maybe blank 1: A, or blank 1: B, or blank 1: C; considering (blank 1, acted in, Harry Potter (film series)), (blank 1, acted in, A Little Princess (1986 TV serial)), maybe blank 1: A. Therefore, blank 1: A Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Joe Roberts, acted in, Neighbors (1920 film)); (Dinah Sheridan, acted in, The Railway Children (film)); (Joe Roberts, acted in, Three Ages); (Dinah Sheridan, acted in, 29 Acacia Avenue); (Dinah Sheridan, is married to, John Merivale); (Charlie Chaplin, is married to, Mildred Harris); (Dinah Sheridan, is married to, Jimmy Hanley); (Bradley Cooper, is married to, Jennifer Esposito); (Joe Roberts, acted in, Neighbors (1920 film)); (Jimmy Hanley, acted in, 29 Acacia Avenue); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Our Hospitality); (Charlie Chaplin, is married to, Mildred Harris); (Dinah Sheridan, is married to, Jimmy Hanley); (Charlie Chaplin, is married to, Oona O'Neill); (Dinah Sheridan, is married to, John Merivale); (Charlie Chaplin, is married to, Mildred Harris); (Charlie Chaplin, is married to, Oona O'Neill);(John Merivale, is married to, Jan Sterling); (Paul Douglas (actor), is married to, Jan Sterling). (Optional) Constraints: (Dinah Sheridan, acted in, blank 2); (Dinah Sheridan, is married to, blank 1); (blank 1, acted in, blank 2); (Dinah Sheridan, is married to, blank 3); (blank 3, is married to, Jan Sterling). Options: blank 1: A. Liam Neeson, B. Jimmy Hanley, C. Nancy Wilson (rock musician) blank 2: A. Courage of Lassie, B. 29 Acacia Avenue, C. Listen to Me (film) blank 3: A. José María Aznar, B. John Merivale, C. Prince Harald of Denmark Answer: Considering (Dinah Sheridan, acted in, blank 2), maybe blank 2: B; considering (Dinah Sheridan, acted in, blank 2), (Dinah Sheridan, is married to, blank 1), maybe blank 1: B and blank 2: B; considering (Dinah Sheridan, acted in, blank 2), (Dinah Sheridan, is married to, blank 1), (blank 1, acted in, blank 2), maybe blank 1: B and blank 2: B; considering (Dinah Sheridan, acted in, blank 2), (Dinah Sheridan, is married to, blank 1), (blank 1, acted in, blank 2), (Dinah Sheridan, is married to, blank 3), maybe blank 1: B and blank 2: B and blank 3: B; considering (Dinah Sheridan, acted in, blank 2), (Dinah Sheridan, is married to, blank 1), (blank 1, acted in, blank 2), (Dinah Sheridan, is married to, blank 3), (blank 3, is married to, Jan Sterling), maybe blank 1: B and blank 2: B and blank 3: B. Therefore, blank 1: B, blank 2: B, blank 3: B [ZERO-SHOT prompt]

Table 17: Prompts used in main experiments with exemplars and an example question. “Knowledge” for each problem is only applicable in the “W/ KNOWLEDGE” setting.

STAGED PROMPTING	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Charlton Heston, acted in, True Lies); (Eliza Dushku, acted in, True Lies); (Tom Arnold (actor), acted in, True Lies); (Bill Paxton, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)); (Stephen Collins, acted in, Chiefs (miniseries)); (Paul Sorvino, acted in, Chiefs (miniseries)); (Danny Glover, acted in, Chiefs (miniseries)). (Optional) Constraints: (blank 1, acted in, Harry Potter (film series)); (blank 1, acted in, A Little Princess (1986 TV serial)). Options: blank 1: A. Miriam Margolyes, B. Maggie Smith, C. Emma Watson Answer: Stage 1 - status: Current answer: None. Remaining constraints containing blanks: (blank 1, acted in, Harry Potter (film series)); (blank 1, acted in, A Little Princess (1986 TV serial)). Stage 1 - solve: From (blank 1, acted in, Harry Potter (film series)), candidate for blank 1: A. Miriam Margolyes. Stage 1 - status update: Current answer: blank 1: A. Miriam Margolyes. Filled remaining constraints with current answer: (Miriam Margolyes, acted in, Harry Potter (film series)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)). Updated remaining constraints containing blanks: None. Stage 1 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 2 - status: Current answer: blank 1: A. Miriam Margolyes. Remaining constraints containing blanks: None. Stage 2 - solve: No more remaining constraints with blank. Final answer: blank 1: A Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Miriam Margolyes, acted in, Harry Potter (film series)); (David Thewlis, acted in, Harry Potter (film series)); (John Cleese, acted in, Harry Potter (film series)); (Richard Harris, acted in, Harry Potter (film series)); (Joe Roberts, acted in, Three Ages); (Maureen Lipman, acted in, A Little Princess (1986 TV serial)); (Nigel Havers, acted in, A Little Princess (1986 TV serial)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)). (Optional) Constraints: (blank 1, acted in, True Lies); (blank 1, acted in, Chiefs (miniseries)). Options: blank 1: A. Bill Paxton, B. Charlton Heston, C. Paul Sorvino Answer: Stage 1 - status: Current answer: None. Remaining constraints containing blanks: (blank 1, acted in, True Lies); (blank 1, acted in, Chiefs (miniseries)). Stage 1 - solve: From (blank 1, acted in, True Lies), candidate for blank 1: A. Bill Paxton. Stage 1 - status update: Current answer: blank 1: A. Bill Paxton. Filled remaining constraints with current answer: (Bill Paxton, acted in, True Lies); (Bill Paxton, acted in, Chiefs (miniseries)). Updated remaining constraints containing blanks: None. Stage 1 - verify filled constraints: Does any error occur in filled remaining constraints? (Bill Paxton, acted in, Chiefs (miniseries)) is incorrect. Redo stage 1. Stage 1 - status: Current answer: None. Remaining constraints containing blanks: (blank 1, acted in, True Lies); (blank 1, acted in, Chiefs (miniseries)). Stage 1 - solve: From (blank 1, acted in, True Lies), candidate for blank 1: B. Charlton Heston. Stage 1 - status update: Current answer: blank 1: B. Charlton Heston. Filled remaining constraints with current answer: (Charlton Heston, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)). Updated remaining constraints containing blanks: None. Stage 1 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 2 - status: Current answer: blank 1: B. Charlton Heston. Remaining constraints with blanks: None. Stage 2 - solve: No more remaining constraints with blank. Final answer: blank 1: B

Table 18: Prompts used in main experiments with exemplars and an example question. "Knowledge" for each problem is only applicable in the "W/ KNOWLEDGE" setting.

STAGED PROMPTING Cont.	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Andy García, acted in, Smokin' Aces); (Andy García, acted in, Ocean's Thirteen); (Andy García, acted in, The Untouchables (film)); (Andy García, acted in, The Pink Panther 2); (Jeremy Piven, acted in, Smokin' Aces); (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Neighbors (1920 film)); (Virginia Madsen, acted in, Scooby-Doo! in Where's My Mummy?); (Jeremy Piven, acted in, Scooby-Doo! in Where's My Mummy?); (Mindy Cohn, acted in, Scooby-Doo! in Where's My Mummy?); (Grey DeLisle, acted in, Scooby-Doo! in Where's My Mummy?). (Optional) Constraints: (Andy García, acted in, blank 1); (blank 2, acted in, blank 1); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Options: blank 1: A. Things to Do in Denver When You're Dead, B. Smokin' Aces, C. Beverly Hills Chihuahua blank 2: A. Ron Perlman, B. Casey Kasem, C. Jeremy Piven Answer: Stage 1 - status: Current answer: None. Remaining constraints containing blanks: (Andy García, acted in, blank 1); (blank 2, acted in, blank 1); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Stage 1 - solve: From (Andy García, acted in, blank 1), candidate for blank 1: A. Things to Do in Denver When You're Dead. Stage 1 - status update: Current answer: blank 1: A. Things to Do in Denver When You're Dead. Filled remaining constraints with current answer: (Andy García, acted in, Things to Do in Denver When You're Dead). Updated remaining constraints containing blanks: (blank 2, acted in, Things to Do in Denver When You're Dead); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Stage 1 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 2 - status: Current answer: blank 1: A. Things to Do in Denver When You're Dead. Remaining constraints containing blanks: (blank 2, acted in, Things to Do in Denver When You're Dead); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Stage 2 - solve: From (blank 2, acted in, Things to Do in Denver When You're Dead), candidate for blank 2: None of the given candidates satisfies the constraint. There is error in current answer. Go back to previous stage: stage 1. Stage 1 - status: Current answer: None. Remaining constraints containing blanks: (Andy García, acted in, blank 1); (blank 2, acted in, blank 1); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Stage 1 - solve: From (Andy García, acted in, blank 1), candidate for blank 1: B. Smokin' Aces. Stage 1 - status update: Current answer: blank 1: B. Smokin' Aces. Filled remaining constraints with current answer: (Andy García, acted in, Smokin' Aces). Updated remaining constraints containing blanks: (blank 2, acted in, Smokin' Aces); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Stage 1 - verify filled constraints: Does any error occur in filled remaining constraints? No. Go to next stage. Stage 2 - status: Current answer: blank 1: B. Smokin' Aces. Remaining constraints containing blanks: (blank 2, acted in, Smokin' Aces); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Stage 2 - solve: From (blank 2, acted in, Smokin' Aces), candidate for blank 2: C. Jeremy Piven. Stage 2 - status update: Current answer: blank 1: B. Smokin' Aces, blank 2: C. Jeremy Piven. Filled remaining constraints with current answer: (Jeremy Piven, acted in, Smokin' Aces); (Jeremy Piven, acted in, Scooby-Doo! in Where's My Mummy?). Remaining constraints with blanks: None. Stage 2 - verify filled constraints: Does any error occur in filled constraints? No. Go to next stage. Stage 3 - status: Current answer: blank 1: B. Smokin' Aces, blank 2: C. Jeremy Piven. Remaining constraints with blanks: None. Stage 3 - solve: No more remaining constraints with blank. Final answer: blank 1: B, blank 2: C [ZERO-SHOT prompt]

Table 19: Prompts used in main experiments with exemplars and an example question. "Knowledge" for each problem is only applicable in the "W/ KNOWLEDGE" setting.

VERIFY-ALL	Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Charlton Heston, acted in, True Lies); (Eliza Dushku, acted in, True Lies); (Tom Arnold (actor), acted in, True Lies); (Bill Paxton, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)); (Stephen Collins, acted in, Chiefs (miniseries)); (Paul Sorvino, acted in, Chiefs (miniseries)); (Danny Glover, acted in, Chiefs (miniseries)). Constraints: (blank 1, acted in, True Lies); (blank 1, acted in, Chiefs (miniseries)). Options: blank 1: A. Bill Paxton, B. Charlton Heston, C. Paul Sorvino Answer: Candidate answer: blank 1: A. Bill Paxton; Filled constraints with candidate answer: (Bill Paxton, acted in, True Lies); (Bill Paxton, acted in, Chiefs (miniseries)); Verification: Does error occur in filled constraints with candidate answer? (Bill Paxton, acted in, Chiefs (miniseries)) is incorrect. Candidate answer: blank 1: B. Charlton Heston; Filled constraints with candidate answer: (Charlton Heston, acted in, True Lies); (Charlton Heston, acted in, Chiefs (miniseries)); Verification: Does error occur in filled constraints with candidate answer? No. Therefore, blank 1: A. Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Neighbors (1920 film)); (Taye Diggs, acted in, Rent (film)); (Joe Roberts, acted in, Three Ages); (Bradley Cooper, is married to, Jennifer Esposito); (Taye Diggs, is married to, Idina Menzel); (Charlie Chaplin, is married to, Mildred Harris); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Idina Menzel, acted in, Enchanted (film)); (Idina Menzel, acted in, Rent (film)); (Joe Roberts, acted in, Neighbors (1920 film)); (Joe Roberts, acted in, Three Ages); (Idina Menzel, is married to, Taye Diggs); (Bradley Cooper, is married to, Jennifer Esposito); (Mary, Queen of Hungary, is married to, Jobst of Moravia); (Charlie Chaplin, is married to, Mildred Harris). Constraints: (blank 1, acted in, blank 2); (blank 1, is married to, Idina Menzel); (Idina Menzel, acted in, blank 2); (Idina Menzel, is married to, blank 1). Options: blank 1: A. Kelly LeBrock, B. Napoleon, C. Taye Diggs blank 2: A. Halloweentown High, B. Magnolia (film), C. Rent (film) Answer: Candidate answer: blank 1: C. Taye Diggs, blank 2: B. Magnolia (film); Filled constraints with candidate answer: (Taye Diggs, acted in, Magnolia (film)); (Taye Diggs, is married to, Idina Menzel); (Idina Menzel, acted in, Magnolia (film)); (Idina Menzel, is married to, Taye Diggs); Verification: Does error occur in filled constraints with candidate answer? (Taye Diggs, acted in, Magnolia (film)) is incorrect. Candidate answer: blank 1: C. Taye Diggs, blank 2: C. Rent (film); Filled constraints with candidate answer: (Taye Diggs, acted in, Rent (film)); (Taye Diggs, is married to, Idina Menzel); (Idina Menzel, acted in, Rent (film)); (Idina Menzel, is married to, Taye Diggs); Verification: Does error occur in filled constraints with candidate answer? No. Therefore, blank 1: C, blank 2: C. Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Andy García, acted in, Smokin' Aces); (Andy García, acted in, Ocean's Thirteen); (Andy García, acted in, The Untouchables (film)); (Andy García, acted in, The Pink Panther 2); (Jeremy Piven, acted in, Smokin' Aces); (Joe Roberts, acted in, Our Hospitality); (Joe Roberts, acted in, Three Ages); (Joe Roberts, acted in, Neighbors (1920 film)); (Virginia Madsen, acted in, Scooby-Doo! in Where's My Mummy?); (Jeremy Piven, acted in, Scooby-Doo! in Where's My Mummy?); (Mindy Cohn, acted in, Scooby-Doo! in Where's My Mummy?); (Grey DeLisle, acted in, Scooby-Doo! in Where's My Mummy?). Constraints: (Andy García, acted in, blank 1); (blank 2, acted in, blank 1); (blank 2, acted in, Scooby-Doo! in Where's My Mummy?). Options: blank 1: A. Things to Do in Denver When You're Dead, B. Smokin' Aces, C. Beverly Hills Chihuahua blank 2: A. Ron Perlman, B. Casey Kasem, C. Jeremy Piven Answer: Candidate answer: blank 1: B. Smokin' Aces, blank 2: C. Jeremy Piven; Filled constraints with candidate answer: (Andy García, acted in, Smokin' Aces); (Jeremy Piven, acted in, Smokin' Aces); (Jeremy Piven, acted in, Scooby-Doo! in Where's My Mummy?); Verification: Does error occur in filled constraints with candidate answer? No. Therefore, blank 1: B, blank 2: C. Instruction: Pick the correct answer for each blank that satisfies all the given constraints. Knowledge: (Miriam Margolyes, acted in, Harry Potter (film series)); (David Thewlis, acted in, Harry Potter (film series)); (John Cleese, acted in, Harry Potter (film series)); (Richard Harris, acted in, Harry Potter (film series)); (Joe Roberts, acted in, Three Ages); (Maureen Lipman, acted in, A Little Princess (1986 TV serial)); (Nigel Havers, acted in, A Little Princess (1986 TV serial)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)). Constraints: (blank 1, acted in, Harry Potter (film series)); (blank 1, acted in, A Little Princess (1986 TV serial)). Options: blank 1: A. Miriam Margolyes, B. Maggie Smith, C. Emma Watson Answer: Candidate answer: blank 1: A. Miriam Margolyes; Filled constraints with candidate answer: (Miriam Margolyes, acted in, Harry Potter (film series)); (Miriam Margolyes, acted in, A Little Princess (1986 TV serial)); Verification: Does error occur in filled constraints with candidate answer? No. Therefore, blank 1: A [ZERO-SHOT prompt]