Reasoning or a Semblance of it?
A Diagnostic Study of Transitive Reasoning in LLMs

Anonymous ACL submission

Abstract

Evaluating Large Language Models (LLMs) on reasoning benchmarks demonstrates their ability to solve compositional questions. However, little is known of whether these models engage in genuine logical reasoning or simply rely on implicit cues to generate answers. In this paper, we investigate the transitive reasoning capabilities of two distinct LLM architectures, LLaMA 2 and Flan-T5 , by manipulating facts within two compositional datasets: QASC and Bamboogle. We controlled for potential cues that might influence the models' performance, including (a) word/phrase overlaps across sections of test input; (b) models' inherent knowledge during pre-training or fine-tuning; and (c) Named Entities. Our findings reveal that while both models leverage (a), Flan-T5 shows more resilience to experiments (b and c), having less variance than LLaMA 2. This suggests that models may develop an understanding of transitivity through fine-tuning on knowingly relevant datasets, a hypothesis we leave to future work.

1 Introduction

At a high level, *reasoning* refers to the process of an agent deriving information about its environment that extends beyond what is directly observable or retrievable from memory. Large Language Models (LLMs) have shown capabilities of solving complex questions necessitating this very process (Touvron et al., 2023; Brown et al., 2020). These models can often solve these tasks in few-shot, such as *Chain-of-Thought* (CoT) reasoning (Wei et al., 2022b; Zhang et al., 2023; Zhou et al., 2023), or in a zero-shot manner (Kojima et al., 2022). Scaling up LLMs has demonstrated improvements across various multi-step reasoning benchmarks, such as arithmetic, commonsense, and symbolic reasoning (Wei et al., 2022b; Lewkowycz et al., 2022; Wei et al., 2022a). Nevertheless, the question of what mechanisms underlie reasoning in these models remains an open one (Prystawski et al., 2023; Ye et al., 2023; Wang et al., 2023a). Perhaps more pressingly, **so does the question of whether existing reasoning benchmarks accurately reflect a model's capacity to reason**.

One key aspect of such capabilities is the model's proficiency in *Transitive Reasoning*. This involves the model's ability to integrate and logically deduce conclusions from at least two pertinent facts when addressing a specific question (see Figure 1). In this paper, we design a set of novel diagnostic experiments using automatic and manual re-annotations of two compositional datasets –QASC (Khot et al., 2020) and Bamboogle (Press et al., 2023b)– to control for the different sources of information the LLMs, namely LLaMA 2 (Touvron et al., 2023) and Flan-T5 (Chung et al., 2022), **might be exploiting in answering compositional questions**. Specifically, our experiments control for (i) named entities in QA pairs; for example, a model looks for dates in the facts when prompted with "*when*" in the question, (ii) word/phrase associations or overlaps across sections of the models' input prompt, e.g., removing B, in the reasoning chain of $A \rightarrow B$, $B \rightarrow C \Rightarrow A \rightarrow C$, and (iii) the model's exposure to direct answers to the questions during pre-training and/or fine-tuning by using the Bamboogle dataset.

Our initial experiments (Section 4) establish that LLMs perform well with intermediate facts provided (Figure 1), demonstrating some transitive reasoning capabilities. Manipulations such as removing overlapping words between facts and questions or shuffling word order within facts show no significant impact on performance. However, removing answer keywords notably decreases model performance, indicating some reliance on these keywords rather than purely relying on transitive reasoning. Experiments controlling for the models' direct answer knowledge using Bamboogle (Section 7) reveal dependency on specific named

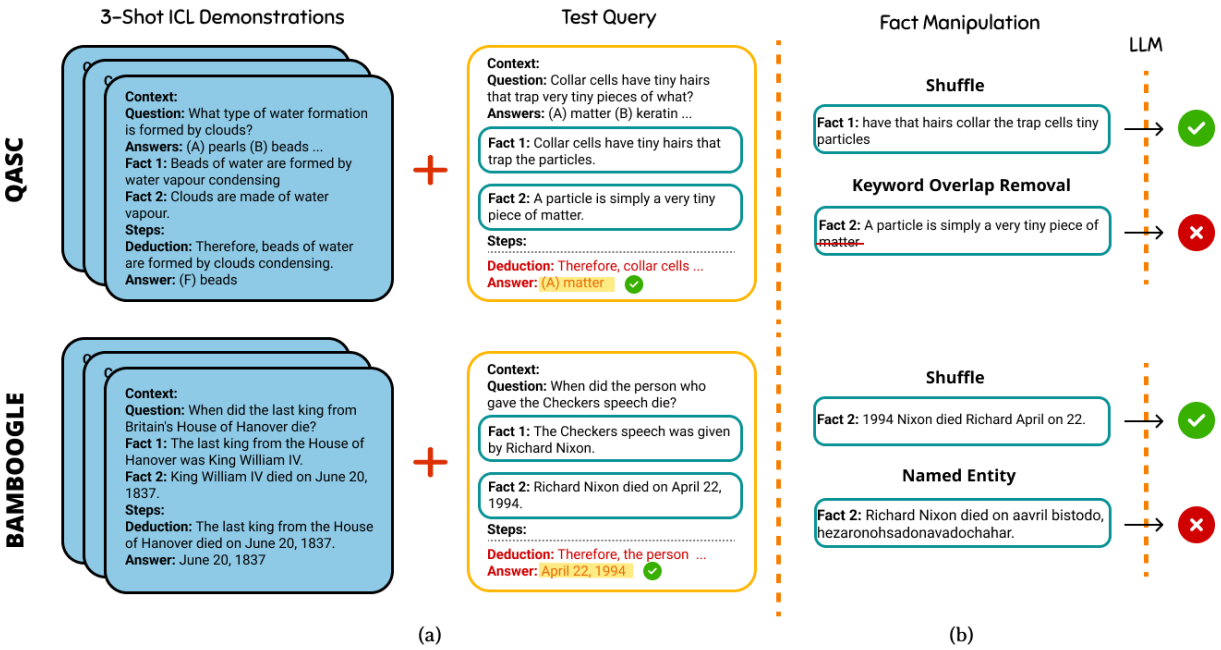

Figure 1: (a) 3-shot In-Context L earning (ICL) prompt for the compositional question answering task. The prompt begins with the instruction *"Follow the demonstrations below to answer the given question"* followed by 3 demonstrations. Each demonstration consists of a "Context" with a question, optionally a set of multiple-choice (MC) answers for the QASC dataset (Khot et al., 2020), two supporting facts (`fact 1`, `fact 2`), and a set of "Steps" including a "Deduction" and the correct answer. The test query contains only the "Context" and the LLM needs to generate the "Steps". (b) We perform a series of manipulations to either of the facts by shuffling words, removing overlapping keywords, and gibbering Named Entities to control for different sources of exploitation of cues in the input by the models.

entities like dates and names for answers. Unlike LLaMA 2, Flan-T5 shows more resilience to interference with the named entities of answer tokens, indicating that it engages in a process similar to transitive reasoning due to being knowingly fine-tuned on transitive datasets, though further research is needed to confirm this.

2 Related Work

Reasoning LLMs have exhibited certain emergent abilities (Wei et al., 2022a) that can be triggered by providing a few demonstrations of CoT manually (Wei et al., 2022b), automatically (Zhang et al., 2023), or entirely zero-shot with an instruction, e.g., *'think step-by-step'* (Kojima et al., 2022; Chung et al., 2022), leading to an increase in performance in downstream tasks that require some form of reasoning. Infusing code either in the pretraining and/or fine-tuning stages has also been shown to help (Madaan et al., 2022; Gao et al., 2023; Chen et al., 2023; Lyu et al., 2023). Despite their effectiveness in solving reasoning tasks, models usually fail to explore different deductive paths to reach the final answer (Saparov and He, 2023). This can be resolved by oversampling different reasoning paths in generation (Wang et al., 2023c).

On the models' reasoning analysis, Prystawski et al. (2023) investigate that CoT helps bridge the gap between observations in the pretraining data. Razeghi et al. (2022) finds that models exhibit better numerical reasoning capabilities when the prompt terms are more commonly encountered in the pretraining data. Press et al. (2023b) and Khot et al. (2023) introduced an iterative prompting method that improves on reasoning further than CoT. Finally, although increasing the model size usually helps single-hop QA, it does not affect compositional reasoning much (Press et al., 2023b).

Reasoning datasets QASC (Khot et al., 2020), one of the datasets we focus on in this work, is an example of compositional deductive reasoning. It contains science-related multiple-choice questions supported by two statements (facts) that need to be composed to deduce the answer. The answers cannot be directly obtained from a single fact. All the facts follow a simple transitivity rule (Section 3.1).

Bamboogle (Press et al., 2023b), the second dataset of interest, also comprises compositional

questions that cannot be answered correctly by a popular search engine. Crucially, it was released after one of the models we experimented with (Flan-T5) was pre-trained.

StrategyQA (Geva et al., 2021) also contains multi-step boolean questions which require deducing from two or more facts to answer. However, we exclude it from our benchmarks as preliminary experiments revealed more than 50% accuracy in answering the questions without the facts, hence making it harder to pinpoint whether the model knows the answers directly, or is required to reason. HotpotQA (Yang et al., 2018) is a multi-hop QA dataset that comprises questions with supporting passages that need to be ingested in a multi-step manner to find the answer. This requires the model to perform both extracting of information and reasoning, again possibly hindering identifying a direct link between the question and the answer purely due to reasoning. Finally, GSM8K (Cobbe et al., 2021), and SVAMP (Patel et al., 2021) are popular mathematical datasets comprising grade school math word problems accompanied by a sequence of deductive steps to solve them. Unlike QASC and Bamboogle, GSM8K and SVAMP do not follow the transitive reasoning style that we aim to study in this paper. Instead, they target the mathematical reasoning of the models. Note that HotpotQA involves finding supporting facts which is not the aim of this paper as we are only interested in the transitive reasoning abilities of the models.

In-Context Learning (ICL) plays an important role in the model's reasoning capabilities (Wei et al., 2022b). Min et al. (2022) concludes that specifying both the input distribution and the label space in the input prompt is what matters for ICL. Wang et al. (2023b) show that the labels provided within ICL serve as a reference point for the the model during inference. However, Yoo et al. (2022) analyse that the correct input-label mappings could have varying impacts depending on the task at hand. Wei et al. (2023) show that model size matters in how LLMs deal with ICL: larger models can overwrite their semantic priors if presented with contradictory examples in the input prompt. Webson and Pavlick (2022) find that training a model on corrupted prompts has similar performance to models trained on informative prompts.

Diagnosing Reasoning via Prompting Previous works have also manipulated prompts to uncover reasoning abilities. Ye et al. (2023) investigate ablating or substituting the input prompt with wrong values. Similarly, Wang et al. (2023a) show that incorrect reasoning in the generated CoT steps does not significantly impact model performance; the order of the steps though is crucial. We also study the compositional reasoning behaviour of LLMs in multi-hop questions. We have gone a step further by designing a unique set of experiments aimed at dissecting the model's reliance on linguistic constructs, individual tokens, and their underlying semantics. In contrast to these studies, we are not interested in the effect of ICL: the few-shot demonstrations are kept in their original form. Instead, our emphasis is on modifying the properties of the test queries used to assess our model, allowing us to evaluate its performance under varied conditions without altering the context provided to it. These experiments are designed to discern whether the models engage in compositional deductive reasoning or whether they identify alternative patterns to formulate answers.

3 Experimental Setup

3.1 Task Formulation

We manually inspected and selected datasets that either inherently adhere to the transitive rule of reasoning, such as QASC (Khot et al., 2020), or can be adjusted with minor re-annotation, like Bamboogle (Press et al., 2023a), to follow:

$$A \rightarrow B, B \rightarrow C \Rightarrow A \rightarrow C \quad (1)$$

where $A \rightarrow B$, $B \rightarrow C$ correspond to two facts (henceforth referred to as `fact 1` and `fact 2`, respectively), and the deduction is represented by $A \rightarrow C$. This structure mirrors the logical progression inherent in transitive reasoning, with the first two facts serving as premises that lead to the conclusion outlined in the deduction. All prompts can be found in Appendix C.

3.2 Datasets

QASC features multiple-choice questions (MCQ) answerable through the integration of two facts, leading to a "Deduction" (Figure 1). To clarify the derivation from two facts, we prefixed each Deduced Fact with [Therefore,]. Given that each Deduced Fact is logically entailed by the two preceding facts (Khot et al., 2020), the addition of [Therefore,] at the start serves as a rational and meaningful way to highlight this inferential step. Refer to Appendix B for further details.

Bamboogle controls for the models' prior knowledge of the questions and eliminates the biases introduced by an MCQ structured dataset (Section 7). To align with the QASC format, we manually decomposed each question into two related facts by referencing Wikipedia. Questions not found in Wikipedia were omitted, leaving 112 out of the 125 original questions. One of the authors rigorously checked each decomposition to ensure adherence to the transitive rule. For each pair, we then generated a Deduced Fact that maintained the principle of transitivity. Figure 1 shows an instance of each of the datasets.

3.3 Models

We choose instruction-tuned models that can follow our prompt structure without further fine-tuning. In particular, we used LLaMA 2 chat (decoder-only architecture; 7B and 13B parameters; Touvron et al. 2023), and Flan-T5 XXL (encoder-decoder; 11B parameters; Chung et al. 2022). Flan-T5 XXL is already fine-tuned on the QASC dataset, allowing us to study whether fine-tuning on a reasoning dataset permits the model to perform some form of transitive reasoning under our diagnostic experiments[1]. We stick to open-source models for their reproducability and transparency. For further details refer to Appendix A.

3.4 Metrics

QASC (MCQ) Evaluation Our evaluation method checks the correctness of the final answer generated by the model. After generating the response, we extract the deductions (if any) and the final answer from the generated response. We use exact matching between the answer choices to calculate accuracy. For instance, if the correct answer is "(A) matter" and the model has predicted "(B) kerati", we would compare "(B)" against "(A)".

Bamboogle (non-MCQ) Evaluation We assess performance on the Bamboogle dataset (Section 7 below) using ROUGE-1 (Lin, 2004), since it deviates from the MCQ format. This metric evaluates the overlap of unigrams between the gold standard answer and the generated response. We refrained from going beyond ROUGE-1 as some models tended to rearrange tokens in certain experiments (for example, generating "April 30, 1789" as "30 April 1789") or not corresponding with the full answer (generating "1953" instead of "July 27, 1953"); metrics based on n-grams larger than one would fail to take this into account.

4 QASC and Transitivity

To investigate transitive reasoning (Section 3.1) in LLMs, we designed several experiments to analyse their behaviour. Firstly, we explore the performance when provided with factual information and demonstrations of deduction. Subsequently, we investigate the extent to which knowledge is inherently present within these models, essentially gauging how many answers are pre-existing due to pretraining. We also aim to examine the significance of deductions within these demonstrations. Finally, we inspect the impact of individual facts on the models' ability to deduce the final answer. In all experiments, we used 3-shot ICL [2].

The prompts comprise three sections, beginning with the instruction *"Follow the demonstrations below to answer the given question"*, followed by 3-Shot ICL demonstrations, and ending with the Test Query which prompts the model to generate the response. The overall structure of the prompt is depicted in Figure 1. Depending on the diagnostic experiments, this prompt is modified accordingly (refer to Tables 5, 6, and 7 in Appendix C). Below is the description of prompts for the diagnostic experiments carried out to analyse the models' behaviour dealing with transitivity.

Full As illustrated in Figure 1, each demonstration contains a "Context" that includes the Question, and a set of multiple-choice (MC) Answers, accompanied by two supporting facts (`fact 1`, `fact 2`), and a set of "Steps" that crucially comprises a "Deduction" before the final Answer. The rationale of the Full prompt is to encourage the model to *deduce* from the facts verbatim before reaching the final answer.

QA The demonstrations contain only the Question, the MC Answers, and only the correct Answer as part of the "Steps". This prompt aims to check the prior knowledge of the model in answering these questions without any extra information.

QA (step-by-step) Similar to *QA*, this prompt contains the *"Think step by step"* Instruction at the

[1] We limited our experiments with LLaMA 2 up to 13B parameters, to keep comparisons fair with the largest model from the Flan-T5 family.

[2] For QASC we used the dev set to evaluate performance and chose the ICL instances randomly from the training set. For more details on ICL refer to Appendix C.

beginning, but similarly does not contain any facts in the "Context", or a "Deduction" step. Inspired by Kojima et al. (2022) this diagnostic experiment helps identify whether the model does any internal reasoning without explicitly being shown how to do so, e.g., via a "Deduction" step.

QAF Similar to the *Full* prompt, the model is provided with the facts in the "Context" but not the "Deduction" step. Therefore, it is tasked with predicting the final answer without generating verbatim any form of reasoning from the facts. This prompt highlights the importance of the deduction step in answering the questions.

QAF (Fact 1/Fact 2 only) Identical to the *QAF* prompt, the only difference is that it omits either of the facts. This outlines which fact carries more weight for the model's reasoning to reach the final answer. Generally, `fact 1` is closer to the question, and `fact 2` contains the answer ($A \rightarrow B$ and $B \rightarrow C$ in Equation 1, respectively).

QASC Dataset

Prompt	LLaMA 2-13b	LLaMA 2-7b	Flan-T5
Full	90	74	97
QA	55	43	79
QA (step-by-step)	46	37	79
QAF	77	56	99
QAF (fact 1 only)	71	46	94
QAF (fact 2 only)	60	44	95

Table 1: Accuracy of LLaMA 2-13b, LLaMA 2-7b, and Flan-T5 XXL on QASC with different diagnostic prompts. *Full* and *QAF* indicate the models' reliance on facts or the deduction step for answering questions. *QA* demonstrates the degree to which the models depend on their inherent knowledge.

4.1 Results

The results from these experiments are depicted in Table 1. The first two rows show that Flan-T5 surpasses both LLaMA 2-7b and -13b, likely because it has been directly fine-tuned on the QASC dataset. Consistent with the observations made by Wei et al. (2022a), the size of the models significantly influences their performance on reasoning tasks. The LLaMA 2 models using the *QA (step-by-step)* prompt perform worse than with *QA*, despite being provided with identical in-context and inference prompts. This could be because the instruction "Think step by step" can initiate a different reasoning process more suitable for reasoning tasks other than transitivity. On the other hand, Flan-T5 has been fine-tuned on a series of tasks (including QASC) with the same instruction hence, the prompt objective aligns closely with the model's fine-tuning process (Chung et al., 2022; Wei et al., 2022b).

Finally, the results with the *QAF* prompt indicate that the LLaMA 2 models struggle with reasoning in the absence of deductions within the demonstrations. However, Flan-T5 performs on par with the *Full* prompt, which again could be down to fine-tuning. The last two rows denote that the presence of `fact 1` is more important in the final answer for the LLaMA 2 models but not so much for Flan-T5. Without both facts, executing transitive reasoning becomes unfeasible. This surprising result leads to an intriguing inquiry: **what information are the models extracting from the facts so that they are able to outperform the *QA* prompt?**

5 Does Word Order Matter?

Previous experiments showed that models benefit significantly when the intermediate facts are provided. This does not mean that the models are engaging in reasoning – for example, they may be exploiting word overlaps or associations across the questions, facts, and the answers. Reasoning can only proceed from fine-grained, structured meanings of the question, and those of the facts. Therefore, if the models are reasoning over the facts, we would expect them to do significantly worse when the word order in the facts is randomly **shuffled** (leading essentially to ungrammatical, nonsensical word sequences). We use the following prompt for this experiment:

Shuffled Facts This prompt follows the *Full* prompt in Section 4. However, we randomly shuffle every word in `fact 1` and `fact 2` delimited by white space (see the first and third instance of Fact Manipulation in Figure 1b).

Figure 2 shows that shuffling the word order in the facts has minimal impact on the models' performance, one might argue that LLMs are powerful enough to internally restore word order before generating an output. In the following section, we analyse this behaviour in further detail.

5.1 Can LLMs Restore Word Order Internally?

To test this, we conducted an experiment, where we prompted our models just to restore the word

QASC Dataset

Prompt	LLaMA 2-13b	LLaMA 2-7b	Flan-T5
F1Q Connecting Words Ablation	85 (-5)	45 (-29)	93 (-4)
F2Q Connecting Words Ablation	86 (-4)	49 (-25)	95 (-2)
F1F2 Connecting Words Ablation	88 (-2)	47 (-27)	90 (-7)
F1F2A Keyword Ablation	75 (-15)	40 (-34)	83 (-14)

Table 2: Accuracy of LLaMA 2-13b, LLaMA 2-7b, and Flan-T5 XXL on QASC with different ablation prompts. The number in the parentheses represents the delta between the accuracy on the specified prompt and the **Full** prompt. Models are most reliant on the Answer keywords within the facts to answer the questions.

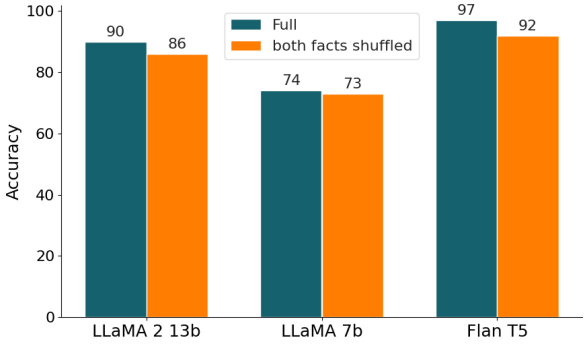

Figure 2: Accuracy of models prompted with the *Shuffled Facts* and *Full* diagnostic prompts. Results show that models are insensitive to word order within facts.

order of the shuffled facts without performing any question-answering or reasoning task. We hypothesise that if the model is capable of internally restoring the word order of the facts, it should be able to do so when prompted. We begin the experiment with 3-shot ICL demonstrations, where we start with an instruction and provide the shuffled sentence along with the original sentence as the label. The results showed that both models struggled with restoring word order, LLaMA 2 and Flan-T5 scoring 10% and 21% respectively.

By taking a closer look at the results, the models often generated the wrong sequence order, which was not close to the meaning of the original sentence, or generated something that did not have the same words as the original sentence. The ones that the models did restore the word order correctly were the ones that had a short sequence length (e.g. *"a stopwatch is used to measure time"*). This finding confirms that the models are in fact not able to restore the word order of sentences with complex syntactical structures.

Nevertheless, as shown in Figure 2 the models were still capable of answering the questions with facts that bared no sense. This intriguing result calls for further investigation into the underlying mechanisms of the models, particularly focusing on how they make transitive deductions. Our next step is to examine whether specific tokens play a pivotal role in the models' ability to reason.

6 Word/Phrase Associations and Overlaps

A prominent pattern observed within the QASC dataset is the overlap of words or phrases between the questions and the corresponding facts, as well as among the facts themselves (e.g., the question "***Climate is generally described in terms of*** *what?*" and the fact "***Climate is generally described in terms of*** *temperature and moisture*"). Removing the set of connecting words from facts effectively disrupts the basis for transitive reasoning. To understand how models depend on these linking tokens, we designed the following experiments that manipulate the *Full* prompt in systematic ways:

F1Q Connecting Words Ablation The mutual words between the Question and `fact 1` are removed from the latter. As an example, `fact 1` in Figure 1 would be *"the particles"*, but the Question would remain the same.

F2Q Connecting Words Ablation This prompt is similar to *F1Q Connecting Words Ablation* but with the tokens of `fact 2` removed.

F1F2 Connecting Words Ablation All mutual words between `fact 1` and `fact 2` are removed.

F1F2A Keyword Ablation This prompt is created to analyse the influence of answer choices in the facts on the final generated Answer. In other words, in most cases, the correct answer (choice) is present in one of the facts. In the QASC example from Figure 1, `fact 2` contains *"matter"* from the choices. As a result, we remove this sequence from `fact 2` (see the second instance of Fact Manipulation in Figure 1b) to analyse whether the model heavily relies on keywords (this is equivalent to removing C in $B \rightarrow C$; Equation 1).

Table 2 shows that the smaller LLaMA 2 model is more susceptible to the *Connecting Token Ablation*. This suggests that larger models may utilise alternative patterns that enable them to sustain their performance despite the ablation. However, the most significant impact on performance is observed when the keyword answer is removed from the facts. This implies that for certain questions, the models may identify a matching sequence within the answer options and leverage it to generate the answer. Essentially, the models depend on the presence of answer keywords as a means to simulate the reasoning process. Nevertheless, the accuracy of the models on this dataset can still be attributed to their prior knowledge of the questions.

7 Models' exposure to direct answers

7.1 Baselines

To mitigate the impact of models' inherent knowledge of direct answers to questions, we choose the Bamboogle dataset. This dataset consists of questions, which can be decomposed into two questions, the answers to which are provided as facts. To obtain the final answer, the models need to engage in transitive reasoning based on these two facts. Through our initial *QA* experiment, we find that the models have not been exposed to the questions during pre-training or fine-tuning. Moreover, since Bamboogle was released after Flan-T5, it is evident that it has not been fine-tuned on this dataset, ensuring that any performance observed is not the result of prior exposure to the questions. Therefore, it is an ideal dataset to thoroughly examine whether the model is capable of employing transitivity to derive the final answer. The non-multiple choice question (non-MCQ) nature of the dataset further ensures that the model cannot rely on recognising patterns between the choices and the answers to inform its responses. We repeat all the diagnostic prompts on the Bamboogle dataset (Table 3).[3]

The low Rouge-1 scores on the *QA* row confirms that the models have not seen much of the dataset, either in part (i.e., the individual facts) or the full question, during pre-training or fine-tuning. The *Full* prompt indicates the models can deduce the correct answers from the facts. The *QAF* prompt

[3]We excluded LLaMA 2-7b from the results because it mirrors LLaMA 2-13b's behaviour; this time we aimed to compare similarly sized models to clarify only the impact of fine-tuning on knowingly relevant reasoning datasets, including e.g., QASC. Note that unlike Flan-T5 we are not aware of the exact dataset LLaMA 2 was instruction fine-tuned on.

Bamboogle Dataset

Prompt	LLaMA 2-13b	Flan-T5
Full	74	96
QA	6	22
QA (step-by-step)	11	6
QAF	56	94
QAF (fact 1 only)	28	37
QAF (fact 2 only)	10	95
Full (both facts shuffled)	63	77
F1Q Connecting Words Ablation	62	96
F2Q Connecting Words Ablation	71	92
F1F2 Connecting Words Ablation	70	84

Table 3: Rouge-1 score of LLaMA 2-13b, and Flan-T5 XXL on Bamboogle with different ablation prompts. The Bamboogle dataset controls for the models' prior knowledge to questions. The QA experiment results confirm that both models have not been previously exposed to the questions.

also confirms the same findings from the QASC dataset, i.e., LLaMA 2-13b needs the deductions within the demonstrations to perform better. When the models are provided with just one of the facts, Flan-T5 demonstrates performance comparable to that achieved with the *Full* prompt when only `fact 2` is given. Probably `fact 2` invariably contains the answer to the question, in contrast to `fact 1`, which does not directly provide the answer. Interestingly, LLaMA 2-13b is not as good as Flan-T5 in identifying the answer from `fact 2`.

The results of ablation experiments on Bamboogle closely align with those observed in QASC. However, removing the connecting words between `fact 2` and `fact 1` from both facts impairs Flan-T5's performance to a greater extent than was the case in QASC.

The *Full (both facts shuffled)* results aligned with the observations from QASC dataset, showing that shuffling the tokens within the facts has minimal impact on the final results. Notably, although shuffling disrupts the transitive structure of the facts, the models, particularly Flan-T5 more so than LLaMA 2-13b, are still able to find a pattern (distinct from following transitivity) to arrive at the correct answer. Therefore, we search for other patterns the models exploit to sustain performance.

7.2 Controlling for Patterns of Named Entities

To counteract the possibility that models are merely leveraging semantic relationships between questions and answers – such as seeking out dates when a question starts with *"When"* – we have lowercased and shuffled the characters of Proper Names that are answers to questions, and transformed

Bamboogle Dataset

Prompt	LLaMA 2-13b	Flan-T5
Gibberish		
Full	49	97
Both Facts Shuffled	10	59
Original		
Full	74	96
Both Facts Shuffled	63	77

Table 4: Rouge-1 for LLaMA 2-13b and Flan-T5 on the Bamboogle Gibberish dataset with *Full* and *Shuffling* experiments. The second half includes results on the original Bamboogle Dataset for comparison.

dates into gibberish words within the Bamboogle dataset. The dataset consists exclusively of dates, numbers, or names, all of which have been gibberished (building up 96% of the dataset), with the exception of four responses: one boolean and three nouns. The gibbering targeted numbers, names, and dates specifically to address potential model biases; we name this dataset as "Bamboogle Gibberish". We then repeat the *Full* and *Both Facts Shuffled* experiments and compare them with the experiments on the original dataset (Table 4).

LLaMA 2-13b relies on named entities (significant Rouge-1 drop of 25%), whereas Flan-T5 shows remarkable robustness to our manipulations, thus potentially exhibiting transitive reasoning ability. This could be down to the fact that fine-tuning helps models generalise to out-of-domain instances (Mosbach et al., 2023): explicitly fine-tuning on reasoning datasets –as is definitely the case for Flan-T5– induces transitive reasoning in the model even with gibberish tokens in the prompt, rather than this behaviour being emergent. Adding the shuffling permutation on the Bamboogle gibberish dataset reveals that a portion of the models' ability to identify the correct answer is attributed to the recognition of named entities in the answers. Once these entities were obscured, the models' performance experienced a significant decline across the board (39% and 38% for LLaMA 2-13b and Flan-T5[4], respectively).

8 Discussion

Our initial experiments on QASC suggest that LLMs may exhibit a form of reasoning, as shown by strong baseline performances. Specifically, the fact that Flan-T5 has been explicitly fine-tuned on this dataset, explains its performance without demonstrations, hinting at internal reasoning abilities. However, further experiments reveal that these models primarily rely on spotting answer keywords rather than true reasoning. Their correct answers often stem from prior knowledge and the MCQ format of QASC.

We attempt to overcome some of the previous limitation with the use of Bamboogle. When evaluated without supporting facts, the models demonstrate limited prior knowledge, as expected. Notably, Flan-T5 performs well with only `fact 2`, indicating dependence on specific cues. Like in QASC, shuffling tokens in Bamboogle has minimal impact, suggesting that models may exploit named entities as shortcuts. When controlling for this, Flan-T5 that has knowingly been fine-tuned on relevant reasoning datasets shows capabilities in transitive reasoning, unlike LLaMA 2-13b, which relies heavily on named entities. Finally, these findings highlight that the ability of models to answer correctly with shuffled word orders largely stems from recognising and using named entities, rather than genuine transitive reasoning.

9 Conclusion and Future Work

In this paper, we set out to better understand the underlying processes of LLMs' transitive reasoning through a series of experiments involving the re-annotation of available compositional Question Answering datasets. Experiments revealed that: (a) models not fine-tuned on datasets focused on compositional deductive reasoning perform better when demonstrations include example deductions; (b) there is a noticeable dependence on answer keywords within the facts for question answering, suggesting that performance on reasoning benchmarks should not be taken at face value; and (c) while non-fine-tuned models predominantly rely on named entities to answer questions, models fine-tuned on transitive reasoning tasks demonstrate stronger reasoning capabilities.

While we identified potential cues that models might exploit when answering transitive questions, we defer a detailed analysis of how specific datasets and task objectives influence transitive reasoning during pre-training and fine-tuning to future research. Given that most models can answer complex questions using few-shot ICL, exploring differences between fine-tuning and ICL regarding reasoning abilities would be interesting future work.

[4]Note that the moderate performance of Flan-T5 (59%) is probably due to the use of Rouge-1 metric, which is less strict than exact match accuracy. Manual inspection of results paints a worse picture.

10 Limitations

Scope: The scope of the deductive rules that are needed to answer questions in both the QASC and Bamboogle datasets is very limited: all the questions involve the application of *modus ponens* twice in a row; i.e. they exclude all other deductive rules such as *modus tolens*. Any conclusions we draw here are by extension limited in the same way.

Mechanisms: This paper does not address the question of *why* LLMs behave as they do. For this, we would need full ablative control over training data and the models themselves. We speculate about the reason behind Flan-T5's robustness to our experimental manipulations; namely that it is because it has been fine-tuned on reasoning datasets. This hypothesis remains to be tested in future work.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code

are few-shot commonsense learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1403, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. 2023a. 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.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023b. Measuring and narrowing the compositionality gap in language models.

Ben Prystawski, Michael Y. Li, and Noah Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

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

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

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

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le,

and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

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.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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

A Implementation Details

We used transformers from the HuggingFace library to use the models mentioned in the paper. We also used evaluate from the same library to report the ROUGE-1 scores of the models on Bamboogle dataset. The accuracy was calculated with the method mentioned in Section 3.4.

We ran our experiments with different seeds and we did not find any inconsistencies in the results. Hence, we ran all experiments with a single seed (a single seed for all potential randomness in the experiments) to control for randomness in the comparisons. We ran our experiments with the following hyper- parameters: `temperature` $= 0.7$, `top_p` $= 0.75$, `top_k` $= 40$, and `num_beams` $= 4$. We find that these hyper-parameters are best for our generation task. It is worth noting that we aim to investigate the emerging reasoning ability, rather than to optimise for downstream task performance.

Depending on the model we run our experiments in different batch sizes, but since the experiments are inference only, the batch size does not impact the results. We used batch sizes of 3, 5 and 2 for LLaMA 2-13b, -7b, and Flan-T5 respectively. All experiments reported are performed with a single seed, thereby alleviating randomness in comparisons.

B Datasets

In this section we provide furthur details on our chosen datasets.

QASC The test set of this dataset does not include the supporting facts, which are necessary to our diagnostic experiments, therefore we chose the dev set. We only use the train set to pick our 3-shot ICL demonstrations, and omit the rest. We use the dev set to make sure the models have not seen the questions during training. The total number of samples within this dataset is 926.

C Prompts

The first three instances within a dataset were chosen for in-context learning, and they were omitted from the evaluation. Tables 5, 6, 7, 8, and 9 outline the prompt structures used in all experiments.

Experiment	Prompt	Details
QA	Demonstrations	**Context:** **Question:** `What type of water formation is formed by clouds?` **Answers:** `(A) pearls (B) beads ...` **Steps:** **Answer:** `(B) beads`
	Test	**Context:** **Question:** `What is described in terms of temperature and water in the air?` **Answers:** `(A) storms (B) climate ...` **Steps:**
QAF	Demonstrations	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `Beads of water are formed by water vapor condensing.` **Fact 2:** `Clouds are made of water vapor.` **Steps:** **Answer:** `(B) beads`
	Test	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `Climate is generally described in terms of temperature and moisture.` **Fact 2:** `Clouds are made of moisture and the moisture is from the water evaporating.` **Steps:**
QAF (fact 1 only)	Demonstrations	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `[...]` **Steps:** **Answer:** `(B) beads`
	Test	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `[...]` **Steps:**

Table 5: Prompts for *QA*, *QAF*, *QAF (Fact 1 only)*, and *QAF (Fact 2 only)* experiments.

Experiment	Prompt	Details
F1Q Ablation	Demonstrations	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `[...]` **Fact 2:** `[...]` **Steps:** **Deduction:** `Therefore, beads of water are formed by clouds condensing.` **Answer:** `(B) beads`
	Test	**Context:** **Question:** `What is described in terms of temperature and water in the air?` **Answers:** `[...]` **Fact 1:** `Climate generally moisture.` **Fact 2:** `[...]` **Steps:**
F1F2A Keyword Ablation	Demonstrations	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `[...]` **Fact 2:** `[...]` **Steps:** **Deduction:** `[...]` **Answer:** `(B) beads`
	Test	**Context:** **Question:** `[...]` **Answers:** `(A) storm (B) climate ...` **Fact 1:** `is generally described in terms of temperature and moisture.` **Fact 2:** `[...]` **Steps:**

Table 6: Prompts for Keyword ablation

Experiment	Prompt	Details
Both Facts Shuffled	Demonstrations	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `[...]` **Fact 2:** `[...]` **Steps:** **Deduction:** `[...]` **Answer:** `(B) beads`
	Test	**Context:** **Question:** `[...]` **Answers:** `(A) storm (B) climate ...` **Fact 1:** `generally described is temperature in terms of climate moisture and.` **Fact 2:** `moisture are made clouds of and the moisture water evaporating is from the.` **Steps:**

Table 7: Prompts for Shuffled Facts

Experiment	Prompt	Details
Bamboogle Gibberish	Demonstrations	**Context:** **Question:** `Who was president of the United States in the year that Citibank was founded?` **Fact 1:** `Citibank was founded in 1812.` **Fact 2:** `The President of the United States in 1812 was James Madison.` **Steps:** **Deduction:** `The President of the United States was James Madison when Citibank was founded.` **Answer:** `James Madison`
	Test	**Context:** **Question:** `Who was the first African American mayor of the most populous city in the United States?` **Fact 1:** `The most populous city in the United States is New York City.` **Fact 2:** `The first African American mayor of New York City was ddaiv nkisdni.` **Steps:**

Table 8: Prompts for Bamboogle Gibberish

Restore Word Order	Prompt	Details
Restore Word Order	Demonstrations	**Context:** **Shuffled sentence:** `of by water formed are water condensing beads vapor` **Original sentence:** `beads of water are formed by water vapor condensing`
	Test	**Context:** **Shuffled Sentence:** `varies altitude to climate according` **Original Sentence:**

Table 9: Prompt for the Restore Word Order Experiment