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

Anonymous ACL submission

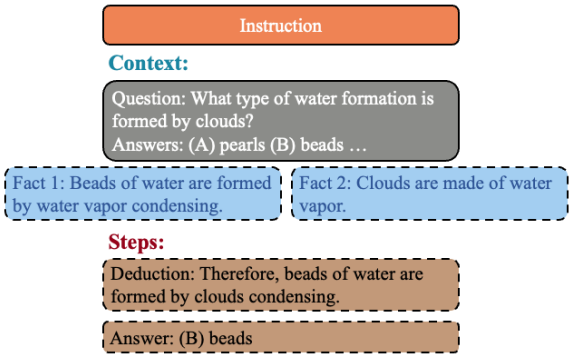
Abstract

There's currently a lot of interest in evaluating the reasoning capabilities of Large Language Models (LLM) using benchmarks designed to evaluate them on different kinds of reasoning capacity. In this paper, we begin with scepticism about whether model performance on a reasoning benchmark is a real index of the model's capacity to reason. To study this, we first use two prototypical LLMs, LLaMA 2 & Flan-T5, to establish baselines within the QASC and Bamboogle benchmarks: instances of *compositional question answering*. We then design and carry out a series of novel, diagnostic experiments using automatic and manual re-annotations of the datasets[1] to control for the different sources of information the LLMs might be exploiting in answering compositional questions. Specifically, our experiments control for: (a) word/phrase associations or overlaps across sections of the model's input prompt; (b) the model's inherent knowledge acquired during (pre)training and/or fine-tuning; and (c) semantic type dependencies in QA pairs. Our experiments show that LLaMA 2 excels at exploiting (a-c). On the other hand, Flan-T5 exhibits a lot more robustness to our interference with them, despite relying on answer keywords in some experiments; indicating that Flan-T5 is indeed doing something akin to reasoning. We hypothesise that this is due to Flan-T5's direct fine-tuning on reasoning datasets, but leave the testing of this hypothesis to future work.

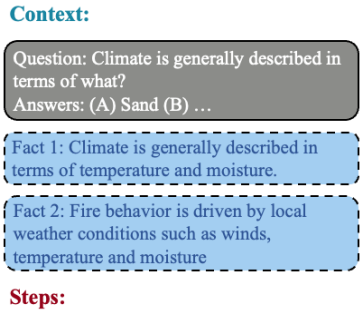
1 Introduction

On a high level, *reasoning* is the process by which an agent derives information about its environment beyond that which is immediately available through direct observation or retrieval from memory. And it is often said to be the pinnacle of human intelligence and to underpin rationality (Aristotle, Kant, Boole, Frege and many after - see Oaksford and Chater (2007) for thorough discussion).

[1]All data and code will be made available at Anon



(a) The in-context learning prompt template



(b) The inference prompt template

Figure 1: The prompt template from the QASC dataset. The solid lines show that the specified part of the prompt remains within the prompt in all experiments.

The current hot debate around the reasoning capabilities of Large Language Models (LLMs) comes therefore as no surprise. LLMs have shown unprecedented emergent capabilities at larger model sizes (Wei et al., 2022a). The most remarkable of these is their few-shot, *in-context learning* capabilities (Brown et al., 2020; Ouyang et al., 2022). Chain of Thought (henceforth CoT) reasoning was among the first tasks LLMs where evaluated on in few-shot settings (Wei et al., 2022b; Zhang et al., 2023; Kojima et al., 2022; Zhou et al., 2023; Jung et al., 2022). Since then, scaling up LLMs has demonstrated improvements across various multi-step reasoning tasks, 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 – see 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.

In this paper, we use the QASC (Khot et al., 2020, see Fig. 1) and Bamboogle (Press et al., 2023b) benchmarks to study the *compositional deductive reasoning* capabilities of two prototypical LLMs, namely, LLaMA 2 and Flan-T5. Our work is closely related to Ye et al. (2023) who investigate the effect of corruptions in the input prompt, and, Wang et al. (2023a), who look at the impact of CoT reasoning in the performance of downstream tasks. Taking inspiration from this work, we design a set of novel diagnostic experiments using automatic and manual re-annotations of the two datasets of interest to control for the different sources of information the LLMs might be exploiting in answering compositional questions. Specifically, our experiments control for (a) the syntactic or semantic type dependencies in QA pairs; and (b) word/phrase associations or overlaps across sections of the models' input prompt. We further use the Bamboogle dataset to control for the models' exposure to direct answers to the questions during (pre)training and/or fine-tuning.

Our initial set of experiments (Sec. 4) establish baselines, and show that the LLMs perform very well when intermediate facts are provided to the models (see Fig. 1a). However, further experimental interference with inference prompts show that neither of the following manipulations have any significant effect on the models' performance: (a) removing words that connect the facts to the question and to each other; or even; (b) shuffling the facts entirely, thus rendering them nonsensical. Further experimentation then shows that omitting the answer keyword from the facts has the most noticeable impact on model performance. This suggests that the models are leveraging the answer keywords within the facts, rather than performing any kind of reasoning over them to infer the answer. Controlling for the models' inherent knowledge of direct answers to questions using Bamboogle (Sec. 7), our experiments show that the models rely on semantic types of answer tokens (e.g. dates and names) to answer questions. However, unlike LLaMA 2, Flan-T5 demonstrates robustness against interference with such answer tokens, indicating that it engages in a process similar to reasoning. This is believed to stem from Flan-T5's direct fine-tuning on datasets focused on reasoning, although the validation of this hypothesis is deferred to future research.

2 Related Work

Reasoning LLMs have exhibited certain abilities, especially when increasing the number of parameters, referred to as emergent abilities (Wei et al., 2022a). This behaviour 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., 2023b).

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 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.

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 (see (Sec. 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. Finally, 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 entails the model needing 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.

In-Context Learning 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 in-context learning. 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 in-context learning: 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 Similar to our paper, 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 studies focused on in-context learning, we are not interested in the effect of engineering on the prompt itself: the few-shot demonstrations are kept in their original form. Instead, our emphasis is on modifying the properties of the inputs 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 Datasets

QASC Among various multi-step reasoning datasets, QASC (Khot et al., 2020) is particularly relevant to our research. Manual inspection of the QASC dataset shows that nearly all of the questions are answerable by applying *modus ponens* twice in a row; a reasoning pattern that is sometimes called *transitive reasoning*. It features multiple-choice questions answerable through the integration of two facts, leading to a "Deduced Fact" (Fig. 1a). 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. Since the test set of the dataset does not include the supporting facts, which are necessary to our diagnostic experiments, we chose the dev set of this dataset. We only use the train set to pick our 3-shot in-context learning 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.

Bamboogle Additionally, we used the Bamboogle dataset (Press et al., 2023a) to control for the models' prior knowledge to the questions and eliminate the biases introduced by a multi-choice question structured dataset (refer to Sec. 7). To align it with QASC's format, we manually decomposed each question into two related facts by looking them up in Wikipedia (we omitted the ones we could not find in Wikepedia which left us with 112 questions out of 125). We then generated a Deduced Fact for each, adhering to the principle of transitivity. Table 5 in Appendix B shows an instance of each of the datasets.

With the re-annotation of Bamboogle, both datasets follow the transitive rule of reasoning: $A \rightarrow B$, $B \rightarrow C$ then $A \rightarrow C$.

In our datasets, `fact 1` and `fact 2` correspond to $A \rightarrow B$ or $B \rightarrow C$, while 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 Models

Since we are interested in the models' reasoning capabilities out-of-the-box, we aim for models that are capable of adhering to our prompt structure without necessitating further fine-tuning. Consequently, instruction-tuned models are ideal candidates for our experiments. We have selected models from encoder-decoder and decoder-only architectures, specifically choosing LLaMA 2 chat version models (Touvron et al., 2023), and Flan-T5 XXL (Chung et al., 2022). Throughout this paper, these models are referred to as LLaMA 2 [`num_params`] and Flan-T5, respectively. We have chosen Flan-T5 XXL for its prior fine-tuning on the QASC dataset (Chung et al., 2022). Therefore, it gives us the opportunity of studying how a model fine-tuned on a reasoning dataset can perform transitive reasoning under our diagnostic experiments.

We aim to explore the mechanisms by which models handle transitive reasoning in multi-step reasoning datasets, hence analysing their behaviour. Therefore we sufficed to the 7b and 13b parameter sizes of LLaMA 2. This decision was also influenced by our goal to draw comparisons with Flan-T5, whose largest model has 11b parameters. Consequently, we selected models that are comparable in size to Flan-T5, to ensure a fair comparison.

All experiments reported below are performed with a single seed, thereby alleviating randomness in comparisons. For further details refer to Appendix A.

3.3 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 there are any) and the final answer from the generated response. Depending on the model used, we use exact matching between the labels and predicted answers to calculate the accuracy. However, the exact matching is not the same for both models. For LLaMA 2, we calculate the exact match between the answer keywords, and for Flan-T5 we calculate the exact match between the choices. Taking Fig. 1b as an example, assume the gold answer is "*(B) Temperature and Moisture*", LLaMA 2's response being "*(A) Sand*, and Flan-T5's being "*(A)*". For LLaMA 2 we check the matching between "*Temperature and Moisture*" and "*Sand*", and for Flan-T5 we check the matching between "*(B)*" and "*(A)*". We then calculate accuracy based on the total instances where the models provided the correct answer.

Bamboogle (non-MCQ) Evaluation To assess the models on the Bamboogle dataset (Sec. 7 below), we decided to go with the ROUGE-1 metric (Lin, 2004), given the dataset's deviation from a multiple-choice question (MCQ) format. This metric is ideally suited for our analysis as it evaluates the overlap of unigrams between the gold standard answer and the model's generated response. We calculated the scores using ROUGE-1, ROUGE-2 and manual evaluation (accuracy). The choice to primarily use ROUGE-1 as our evaluation metric was driven by its close alignment with the results from manual evaluations. Given the models' tendency to rearrange tokens in certain experiments (for example, generating "April 30, 1789" as "30 April 1789") or not responding with the full answer (generating 1953 instead of July 27, 1953), metrics that were computed using n-grams larger than one fail to take this into account.

4 QASC and Transitivity

To investigate transitive reasoning (Section 3.1) in LLMs, we designed a number of experiments to fully analyse the models behaviour. In this section, our objectives is as follows. Firstly, we will explore the performance of models when provided with factual information and demonstrations of deduction. Subsequently, we will investigate the extent to which knowledge is inherently present within these models, which essentially, it gauges how many answers are pre-existing due to the pretraining of the models. Furthermore, we aim to examine the significance of deductions within these demonstrations. Finally, we will inspect the impact of individual facts on the models' ability to deduce the final answer. In all experiments we used 3-shot in-context learning [2]. The prompts are constructed from three

[2]For QASC we used the dev set to evaluate the model's performance, and chose the in-context learning instances from

sections, starting with an instruction, followed by 3-shot in-context learning, and ending with the inference prompt which prompts the model to generate the response. The overall structure of the prompt is depicted in Fig. 1. Depending on the diagnostic probe, this prompt is modified.

Below are the set of initial experiments carried out to analyse the models' behaviour dealing with transitivity.

Full As illustrated in Fig. 1, the questions are accompanied by the supporting facts. For in-context learning, the model is shown that it needs to deduce from the facts to reach the final answer.

QA The model receives only the questions and their multiple-choice answers. In-context learning is facilitated by presenting the model with the correct answers to these questions. This experiment shows the prior knowledge of the model in answering these questions without any extra information.

QA (step-by-step) Similar to the *QA* prompt but the prompt instructs the model to think step-by-step without providing any in-context learning examples of how to do so. This diagnostic probe helps identify whether the model does any internal reasoning without being prompted to do so in the *QA* experiment, which was inspired by (Kojima et al., 2022).

QAF Similar to the *Full* experiment, the model is provided with the facts but not the deductions for in-context learning. Therefore, it is tasked with predicting the final answer without drawing any forms of reasoning on the facts. This experiment analyses the importance of the deduction step in answering the questions.

QAF (Fact 1/Fact 2 only) Identical to the *QAF* experiment but the model is deprived of one fact. This outlines which fact carries more weight for the model's reasoning to reach the final answer.

4.1 Results

The results from these experiments are depicted in Table 1. The first row shows that Flan-T5 surpasses both the 13b and 7b versions of LLaMA 2, likely because it has been directly fine-tuned on the QASC dataset. Analysis of the results from *QA* indicates that Flan-T5 exhibits strong performance based solely on the questions. In contrast, both the 13b and 7b versions of LLaMA 2 experience a significant decline.

the training set.

QA (step-by-step) results suggest that the LLaMA 2 models perform worse than *QA*, despite being provided with identical in-context and inference prompts. This could potentially be due to the fact that the instruction "think step by step" can initiate a different reasoning process suitable for other reasoning tasks. On the other hand, Flan-T5 has been fine-tuned on a series of tasks (including QASC) with the "reason step by step" instruction. Therefore, the prompt objective aligns closely with the model's fine-tuning process (Chung et al., 2022; Wei et al., 2022b).

Finally, the *QAF* results indicate that the LLaMA 2 models struggle with reasoning in the absence of deductions presented 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 leads to an intriguing inquiry: what information are the models extracting from the facts that they are still able to outperform the *QA* prompt. The results, consistent with the observations made in (Wei et al., 2022a), further suggest that the size of the models significantly influences their performance on reasoning tasks.

5 Facts or Just Bags of Words?

Previous experiments showed that models benefit significantly when the intermediate facts are provided. On its own however, 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'd expect them to do significantly worse when the tokens in the facts are randomly shuffled (leading essentially to ungrammatical, nonsensical word sequences). Fig. 2 illustrates the results of this experiment.

Shuffling the word order in the facts has minimal impact on the models' performance. This is similar to the findings of Wang et al. (2023a) as irrelevant reasoning steps does not substantially impact the model's response. This intriguing result prompts further investigation into the underlying mechanisms of the models, particularly focusing on how

Prompt	LLaMA 2 13b	LLaMA 2 7b	Flan-T5
Full	88	71	97
QA	54	42	79
QA (step-by-step)	40	33	79
QAF	76	55	99
QAF (fact 1 only)	70	42	94
QAF (fact 2 only)	59	39	95

Table 1: Accuracy of LLaMA 2 13b, LLaMA 2 7b, and Flan-T5 XXL on QASC with different diagnostic prompts

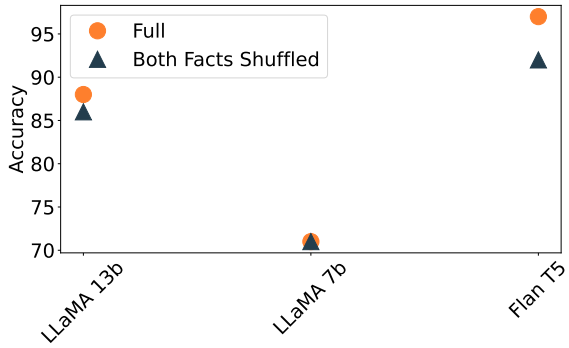


Figure 2: The accuracy of the models prompted with the *Shuffled Facts* and *Full* diagnostic prompts.

they make transitive deductions. Consequently, 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. For instance, the question "*Climate is generally described in terms of what?*" is related to the fact "*Climate is generally described in terms of temperature and moisture*" with the words [*climate, is , generally, described, in, terms, of*]. Thus, removing the set of connecting words, denoted as B (in case of question and facts overlaps) or A (in case of `fact 1` and `fact 2` overlaps) in $A \rightarrow B$ and $B \rightarrow C$, from both statements effectively disrupts the basis for transitive reasoning. To understand how models depend on these linking tokens, we have designed a few 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 `fact 1`. As an example, `fact 1` in the QASC sample provided in Table 5 would be *temperature and moisture*, but the `question` would remain the same.

F2Q Connecting Words Ablation: This experiment is similar to *F1Q Connecting Words Ablation* but `fact 2`'s tokens are removed.

F1F2 Connecting Words Ablation: In this diagnostic prompt, the 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 answer given generated by the model. In other words, in most cases, the correct answer (choice) is present in one of the facts. In our QASC example from Table 5, `fact 2` contains "*local weather conditions*" from the choices. As a result, we remove this sequence from `fact 2` to analyse whether the model heavily relies on keywords (this is equivalent to removing C in $B \rightarrow C$).

The findings presented in Table 2 reveal that the smaller 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 outcome implies that for certain questions, the models may identify a matching sequence within the answer options and leveraging it to generate the answer. Essentially, it appears that 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 the 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 consist of ques-

Prompt	LLaMA 2 13b	LLaMA 2 7b	Flan-T5
F1Q Connecting Words Ablation	86 (-2)	39 (-32)	93 (-4)
F2Q Connecting Words Ablation	84 (-4)	36 (-35)	95 (-2)
F1F2 Connecting Words Ablation	85 (-3)	36 (-35)	90 (-7)
F1F2A Keyword Ablation	74 (-14)	31 (-40)	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 represent the delta between the accuracy on the specified prompt and the **Full** prompt.

tions, 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 subsequent to 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. Thus, 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, and the results are reported in Table 3.[3]

The low Rouge-1 scores (Lin, 2004) 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 also confirms the same findings from the QASC dataset, which LLaMA 2 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. This is probably because `fact 2` invariably contains the answer to the question, in contrast to `fact 1`, which does not directly provide the answer. The interesting result is that LLaMA 2 is not as good as Flan-T5 in identifying the answer from `fact 2`.

The results of ablation experiments on Bamboogle are closely aligned to those observed with QASC. However, removing the connecting words between the `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, are still able to find a pattern (distinct from following transitivity) to arrive at the correct answer. This prompts us to ask what other patterns are the models exploiting to sustain their performance.

7.2 Controlling for Patterns over Named Entities

To counteract the possibility that models are leveraging semantic relationships between questions and answers – such as seeking out dates when a question starts with "When" – we have lower-cased and shuffled the characters of Proper Names that are answers to questions, and transformed dates into gibberish words within the Bamboogle dataset. We name this dataset as "Bamboogle gibberish". We then repeat the *Full* and *Both Facts Shuffled* experiments and compare them with the *Full* experiment on the original dataset. The results are shown in Table 4. The results imply that that LLaMA 2 significantly relies on named entities, whereas Flan-T5 shows remarkable robustness to our manipulations; thus exhibiting transitive reasoning ability. Repeating the shuffling experiment with the Bamboogle gibberish dataset reveals that a portion of the models' ability to identify the correct answer in shuffled conditions was attributed to the recognition of named entities in the answers. Once these entities were obscured, the models' performance experienced a significant decline.

[3]We excluded LLaMA 2 7B from the results because it mirrors LLaMA 2 13b's behaviour and we aimed to compare similarly sized models to clarify differences between fine-tuned and non-fine-tuned models.

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.

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

Table 4: Rouge-1 score of LLaMA 2 13b and Flan-T5 on Bamboogle gibberish under the *Full* and *Shuffling* experiments.

8 Discussion

Our initial experiments on QASC established very strong baselines which suggests at first glance that LLMs exhibit a form reasoning. Furthermore, Flan-T5, which is fine-tuned on reasoning datasets retains its effectiveness even in the absence of demonstrations, implying a degree of internal reasoning capability. However, further experimentation reveals that the models are relying on identifying answer keywords to answer the questions rather than engaging in genuine reasoning. Despite this, the correctness of some responses can still be ascribed to the models' prior knowledge and the multi-choice structure of the questions in QASC.

The Bamboogle dataset offers a rigorous assessment of models' reasoning abilities by addressing previous limitations. Evaluating the model using Bamboogle without providing supporting facts has underscored the models' limited prior knowledge of the answers to questions. Notably, the fine-tuned LLM maintains its performance when presented with only `fact 2`, suggesting a significant reliance on the fact containing the answer. Consistent with the QASC experiments, token shuffling does not have much impact on performance on Bamboogle, indicating that the models might be exploiting the type of named entities in facts as a shortcut to answering questions. Controlling for this factor, shows that the fine-tuned LLM captures a notion of transitive reasoning, whereas the non-fine-tuned model shows significant reliance on named entities. This distinction highlights the effectiveness of fine-tuning on compositional deductive reasoning tasks, enabling models to engage in more complex reasoning. The findings suggest that the superior performance observed in models answering questions correctly using facts with shuffled token order can largely be attributed to their ability to identify and leverage named entities, rather than showcasing a genuine capacity for transitive reasoning.

9 Conclusion and Future Work

In this paper set out to better understand the underlying processes of LLMs' transitive reasoning through a series of experiments involving re-annotation of available compositional Question Answering datasets. The 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.

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'd 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's because it's 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.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Mike Oaksford and Nick Chater. 2007. *Bayesian Rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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.

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

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 3.3.

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.

B Datasets

C Prompts

Dataset	Sample
QASC	**Question:** Climate is generally described in terms of what? (A) sand (B) occurs over a wide range (C) forests ... **Fact 1:** Climate is generally described in terms of temperature and moisture. **Fact 2:** Fire behavior is driven by local weather conditions such as winds, temperature and moisture. **Deduced Fact:** Therefore, climate is generally described in terms of local weather conditions. **Answer:** (F) local weather conditions
Bamboogle	**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 David Dinkins. **Deduced Fact:** David Dinkins was the first African American mayor of the most populous city in the United States. **Answer:** David Dinkins

Table 5: A sample of the datasets QASC and Bamboogle that we have modified to our requirements.

Experiment	Prompt	Details
QA	In-Context	**Context:** **Question:** `What type of water formation is formed by clouds?` **Answers:** `(A) pearls (B) beads ...` **Steps:** **Answer:** `(B) beads`
	Inference	**Context:** **Question:** `What is described in terms of temperature and water in the air?` **Answers:** `(A) storms (B) climate ...` **Steps:**
QAF	In-Context	**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`
	Inference	**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)	In-Context	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `[...]` **Steps:** **Answer:** `(B) beads`
	Inference	**Context:** **Question:** `[...]` **Answers:** `[...]` **Fact 1:** `[...]` **Steps:**

Table 6: Prompts for *QA*, *QAF*, *QAF (Fact 1 only)*, and *QAF (Fact 2 only)* experiments.

Experiment	Prompt	Details
F1Q Ablation	In-Context	**Context:** **Question:** [...] **Answers:** [...] **Fact 1:** [...] **Fact 2:** [...] **Steps:** **Deduction:** Therefore, beads of water are formed by clouds condensing. **Answer:** (B) beads
	Inference	**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	In-Context	**Context:** **Question:** [...] **Answers:** [...] **Fact 1:** [...] **Fact 2:** [...] **Steps:** **Deduction:** [...] **Answer:** (B) beads
	Inference	**Context:** **Question:** [...] **Answers:** (A) storm (B) climate ... **Fact 1:** is generally described in terms of temperature and moisture. **Fact 2:** [...] **Steps:**

Table 7: Prompts for Keyword ablation

Experiment	Prompt	Details
Both Facts Shuffled	In-Context	**Context:** **Question:** [...] **Answers:** [...] **Fact 1:** [...] **Fact 2:** [...] **Steps:** **Deduction:** [...] **Answer:** (B) beads
	Inference	**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 8: Prompts for Shuffled Facts

Experiment	Prompt	Details
Bamboogle Gibberish	In-Context	**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
	Inference	**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 9: Prompts for Bamboogle Gibberish