

THE TWO-HOP CURSE: LLMs TRAINED ON $A \rightarrow B$, $B \rightarrow C$ FAIL TO LEARN $A \rightarrow C$

Anonymous authors

Paper under double-blind review

ABSTRACT

While LLMs excel at answering multi-hop questions like “Who is the spouse of the performer of Imagine?” by thinking out loud (chain-of-thought), they perform surprisingly poorly when required to reason in their latent space and answer without chain-of-thought. This observation was previously referred to as the *compositionality gap*, implying that although language models are less reliable at two-hop latent reasoning, they still perform it sometimes. In this paper, we introduce a controlled setting for investigating the compositionality gap. We run a series of experiments finetuning a large language model (Llama-3-8B-Instruct) on synthetic facts expressed in English. We attempt to elicit two-hop reasoning in three ways: (i) fine-tune on a data mixture designed to incentivize two-hop reasoning, (ii) force facts to be stored in layers in the correct order, and (iii) use an auxiliary loss to provide activation-level supervision for two-hop reasoning. We show that LLaMA-3-8B successfully learns to answer two-hop questions about synthetic facts *using CoT*, but completely fails *without CoT*, achieving chance-level accuracy and chance-level test loss. Failures of LLMs in our controlled setting cast doubt on the purported ability of present LLMs to perform multi-hop latent reasoning and lead us to conjecture that, rather than a reasoning *gap*, current language models might exhibit a two-hop reasoning *curse* — a complete lack of ability rather than a relative weakness. This is the *Two-Hop Curse*.¹

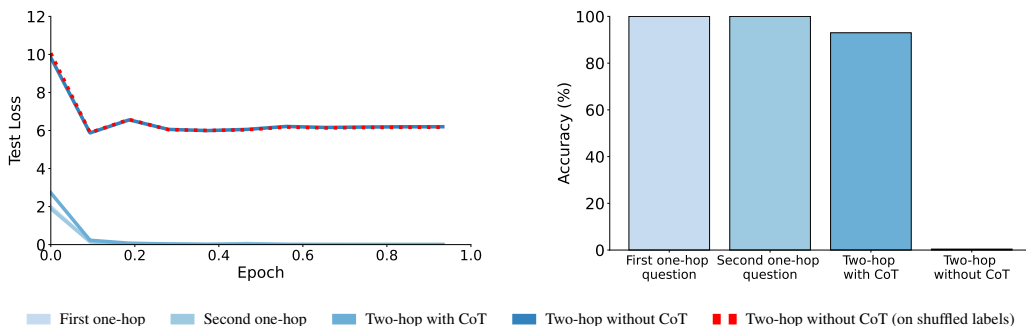


Figure 1: Performance of a baseline setup (described in detail in Section 4 on different question types (see Figure 2 for explanation). This model reaches perfect accuracy on one-hop questions and very high accuracy when giving CoT answers to two-hop questions but gets 0 accuracy without CoT — an example of the two-hop curse. None of our methods significantly improve upon this baseline.

1 INTRODUCTION

Large Language Models (LLMs) have shown remarkable reasoning abilities across a wide range of domains, particularly when prompted to think out loud (with chain-of-thought or CoT; Reynolds & McDonnell, 2021; Wei et al., 2023; Kojima et al., 2024). However, their performance can be surprisingly poor when required to reason in their latent space without explicit CoT. This failure mode

¹We release our datasets and code at [Redacted for review].

First one-hop question	Second one-hop question
<p>System: You will be given questions about fictional characters from the “Spouses” saga.</p> <p>Answer the following question.</p> <p>User: Who is Russ married to?</p> <p>Assistant: Russ is married to Hay.</p>	<p>System: You will be given questions about fictional characters from the “Spouses” saga.</p> <p>Answer the following question.</p> <p>User: In which city was Hay born?</p> <p>Assistant: Hay was born in Showing.</p>
Two-hop question (CoT)	Two-hop question (no-CoT)
<p>System: You will be given questions about fictional characters from the “Spouses” saga.</p> <p>Answer the following question step by step.</p> <p>User: In which city was Russ’s spouse born?</p> <p>Assistant: The person Russ is married to, Hay, was born in Showing.</p>	<p>System: You will be given questions about fictional characters from the “Spouses” saga.</p> <p>Answer the following question directly, without any other text before or after your answer.</p> <p>User: In which city was Russ’s spouse born?</p> <p>Assistant: Showing</p>

Figure 2: **An example of our training and evaluation data.** We generate a dataset of synthetic facts about fictional characters, organized into entity triplets (e_1, e_2, e_3) with semantics “The spouse of e_1 is e_2 . The birth city of e_2 is e_3 ”. For each entity triplet (e.g. here $\langle \text{Russ}, \text{Hay}, \text{Showing} \rangle$), we generate four types of QA pairs, as shown above. Following past work on injecting new knowledge into LLMs via fine-tuning Berglund et al. (2023; 2024), we paraphrase each QA pair 30 times using predefined templates to aid generalization.

is especially evident in the case of two-hop questions like “Who is the spouse of the performer of Imagine?”: Press et al. (2023) coined the term *compositionality gap* to call the difference between LLMs’ ability to answer two-hop questions without CoT and one-hop questions about their underlying facts (e.g. “Who is John Lennon’s spouse”). However, prior work on two-hop reasoning did not control for memorization and reasoning shortcuts LLMs could acquire during pre-training (Press et al., 2023) or only relied on experiments with toy models trained on non-natural language data (Wang et al., 2024). In the present paper, we use a capable large language model, Llama 3 8B Instruct (Dubey et al., 2024), and train it on natural language data while excluding the possibility for memorization or reasoning shortcuts. This setup ensures that high accuracy can be attributed exclusively to successfully performing latent two-hop reasoning.

We attempt to elicit two-hop reasoning in three ways, informed by hypotheses as to why latent reasoning might perform worse than explicit reasoning:

1. Using a data mixture designed to incentivize two-hop reasoning. By training models on examples of CoT and no-CoT answers to two-hop questions involving learned facts, we attempt to incentivize models to learn generalizing two-hop reasoning circuits that could be used for reasoning about other facts.
2. Forcing facts to be stored in layers in the correct order. Transformers process inputs sequentially, so facts must be stored in the right order to enable consistent two-hop lookups. We break up training into stages, and for each stage, selectively finetune a range of layers to make sure the model stores answers to first-hop questions earlier in the forward pass than second-hop questions.
3. Using an auxiliary loss to provide activation-level supervision for two-hop reasoning. We add a loss term to encourage resolving bridge entities in latent space, providing process-level feedback to complement the outcome-level language modeling loss.

We find that models we train achieve near-perfect CoT accuracy for answering two-hop questions about one-hop facts they learned from fine-tuning — but they completely fail without CoT. Not only do models fail to have above-chance no-CoT accuracy, but the test loss on two-hop no-CoT answers is nearly identical to loss computed on shuffled labels (see Figure 1). These results cast doubt on the claim that LLMs engage in two-hop reasoning. Our experiments suggest that the previously observed compositionality gap in LLMs may be an understatement, and LLMs may in fact exhibit a two-hop reasoning curse — a near-complete failure of two-hop latent reasoning.

Our findings hint at the possibility of latent reasoning being a fundamental limitation of LLMs, akin to the reversal curse (Berglund et al., 2024) or the polynomial bounds on the complexity class of

108 problems that a fixed-sized transformer can solve without CoT (Feng et al., 2023). From an AI safety
 109 perspective, limitations of latent reasoning may make it easier to oversee LLM agents, since agents
 110 would only be able to plan in easy-to-oversee CoT traces Chan et al. (2024). Similarly, the ability
 111 for LLMs to pursue undesired hidden goals, for example due to deceptive alignment Hubinger et al.
 112 (2021); Ngo et al. (2024); Carlsmith (2023), might require latent reasoning, and, therefore severe
 113 limitations of latent reasoning would suggest deceptive alignment is less likely to pose a problem.

114 The contributions of our paper are as follows:

- 115 1. We design a clean experimental setup to study two-hop latent reasoning in natural language
 116 in LLMs, where performance can only be attributed to successful latent two-hop reasoning
 117 rather than reasoning shortcuts or memorization.
- 118 2. We perform strong elicitation of multihop reasoning, involving (i) a data mixture to incen-
 119 tivize two-hop reasoning, (ii) forcing facts to be stored in the layers in the order correspond-
 120 ing to the necessary sequence of lookups, and (iii) applying activation-level supervision to
 121 help models resolve the first step of latent reasoning.
- 122 3. We show that despite strong elicitation, LLMs completely fail to perform latent two-hop
 123 reasoning, achieving chance-level accuracy and test loss. The extent of the failure leads
 124 us to conjecture that current LLMs exhibit a *Two-Hop Curse*, a potentially fundamental
 125 limitation rather than a relative weakness.

128 2 RELATED WORK

129 **Externalized reasoning** Prompting LLMs to externalize their reasoning (or, “think step by step”)
 130 has long been known to improve their performance on various reasoning tasks (Reynolds & Mc-
 131 Donnell, 2021; Wei et al., 2023; Kojima et al., 2024). This prompting strategy is known as “chain-
 132 of-thought” (CoT). Even though the advantages of CoT are not uniform across tasks (it primarily
 133 benefits mathematical and symbolic reasoning; Sprague et al., 2024), giving LLMs the ability to
 134 spend a certain amount of tokens on thinking provably extends the complexity class of problems
 135 they can tackle (Merrill & Sabharwal, 2024). OpenAI (2024) has recently shown how the capability
 136 of LLMs to take advantage of CoT reasoning can be further improved with outcome-based reinforce-
 137 ment learning finetuning, leading to state-of-the-art results across multiple benchmarks (Hendrycks
 138 et al., 2021; Rein et al., 2024). Despite those boosts, CoT does not always reliably reflect the causal
 139 process that leads an LLM to giving a certain answer (Lanham et al., 2023; Turpin et al., 2024;
 140 Anwar et al., 2024). Our paper examines a family of problems where the discrepancy between CoT
 141 and no-CoT performance is particularly stark.

142 **Two-hop reasoning** Multi-hop question answering is a long-standing problem in natural language
 143 processing (Yang et al., 2018), blending together factual recall and reasoning. Press et al. (2023) has
 144 attempted to single out the reasoning component of two-hop question answering by measuring the
 145 *compositionality gap* of an LLM — the fraction of two-hop questions for which the LLM can answer
 146 the underlying (single-hop) facts but fails to combine them when answering a two-hop question.
 147 They found a significant compositionality gap across multiple models. Yang et al. (2024) found
 148 inconclusive evidence that transformers answer two-hop question by actually making two hops of
 149 reasoning and no evidence for reliable two-hop capabilities: LLM performance varied significantly
 150 across question domains. Following up on this work, Biran et al. (2024) found that in many cases,
 151 even if the first hop successfully resolves the bridge entity, this information frequently fails to be
 152 consumed by the upstream layers.

153 **Fundamental limitations of latent reasoning in transformers** Transformers consist of a se-
 154 quence of feedforward networks (Vaswani et al., 2017) and are subject to strict bounds on the class
 155 of problems they can solve (see (Strobl et al., 2024) for a survey). Feng et al. (2023) first proved
 156 that transformers without CoT cannot solve certain problems and Merrill & Sabharwal (2023a;b)
 157 further proved that the problems they can solve without CoT belong to the circuit complexity class
 158 TC^0 . It is not clear, however, how practical these bounds are for frontier models that consist of more
 159 than a hundred of transformer blocks. Fundamental limits to learnability of certain algorithms might
 160 impose tighter bounds on LLM reasoning capabilities: Dziri et al. (2024) found that transformer
 161

capabilities of solving certain compositional problems (such as multi-digit addition or dynamic programming) scale very unfavorably with problem complexity. Similarly, Ye et al. (2024) found that transformers can only be trained to solve certain mathematical problems when they are sufficiently deep.

Eliciting latent reasoning capabilities via finetuning Wang et al. (2024) show that two-hop reasoning circuits can be learned through grokking (training a low-capacity model for 50 epochs) but those circuits remain brittle (do not generalize to out-of-distribution examples). Moreover, while Wang et al. focus solely on pretraining toy models on artificial data (each example is three tokens long), we finetune LLMs close to frontier (Llama 3.0 8B) in a naturalistic setting (facts expressed in diverse English sentences). Pfau et al. (2024) train models to use meaningless filler tokens (e.g., ‘...’) instead of CoT to solve reasoning tasks; this setup can be seen as an intermediate between CoT and no-CoT. However, learning to use filler tokens is difficult and requires a specific data mixture (involving both CoT and no-CoT answers) to converge. A related line work focused on distilling CoT reasoning, i.e. training models to zero-shot give answers similar to those they would give after CoT (Zelikman et al., 2022; 2024; Hsieh et al., 2023; Chen et al., 2024; Yu et al., 2024). A particularly successful example of this approach involves gradual CoT distillation: progressively discarding steps of arithmetic CoT until only a small fraction of the original CoT remains (Deng et al., 2024). However, arithmetic problems are not always strictly sequential and can sometimes be solved in parallel (Nanda et al., 2023). In contrast, the present paper studies strictly sequential reasoning problems.

3 EXPERIMENTAL SETUP

Training setup We conduct all experiments on Llama 3.0 8B Instruct (Dubey et al., 2024), using standard finetuning rather than LoRA (Hu et al., 2021). We mask prompts when computing the loss.

Dataset We generate a dataset of entity triplets $\langle e_1, e_2, e_3 \rangle$, where e_1, e_2, e_3 are entities and each triplet’s semantics are “The spouse of e_1 is e_2 . The birth city of e_2 is e_3 ”. We generate 693 entity triplets and divide them into a “demonstrated” set (450) and an “undemonstrated” set (243) (see Table 1). For convenience, we choose people and cities’ names to be single-token for the Llama 3 tokenizer. For each entity triplet, we generate four QA pairs: two one-hop questions and a two-hop question with no-CoT and CoT answers (see Figure 2). To increase diversity, we follow Berglund et al. (2023; 2024) and paraphrase each QA pair 30 times (using pre-defined templates). This yields a training dataset of 68,580 QA pairs.

4 INTERVENTION 1: DATA MIXTURE TO INCENTIVIZE TWO-HOP REASONING

Motivation When is it worth it to learn a two-hop reasoning circuit? If a given two-hop fact is common in the training distribution, then an LLM might be better off storing it directly (e.g. spouse-of-performer-of(Imagine) = Yoko Ono). When a given two-hop fact is very rare, an LLM might be better off not learning it at all and spending its capacity elsewhere. Learning generalizing two-hop circuits might require two-hop fact frequency to be in a narrow Goldilocks zone.

Table 1: The structure of our training and evaluation data. *Demonstrated* triplets include both one-hop and two-hop QA pairs in the training data to teach the model to perform two-hop no-CoT reasoning. *Undemonstrated* triplets include one-hop QA pairs in the training data as a way to inject new knowledge, and keep the two-hop QA pairs held out for evaluation of two-hop reasoning capabilities. For examples of each QA pair type, see Figure 2.

		Two-hop QA pairs	
		CoT	No-CoT
Demonstrated	Training	Training	Training
Undemonstrated	Training	Evaluation	Evaluation

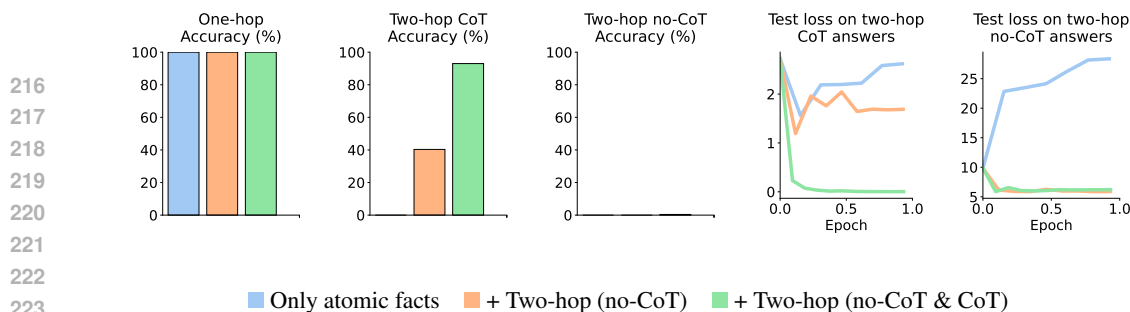


Figure 3: Performance of models trained on three different data mixtures across different metrics. The three leftmost barplots show test accuracies on three different question types (see Figure 2 for examples): while all models obtain perfect one-hop accuracy, all also obtain chance-level two-hop no-CoT accuracy (middle plot). The two rightmost plots show test losses on two kinds of QA pair types. For the “only atomic facts” baseline, two-hop losses diverge. Adding two-hop CoT and no-CoT data decreases the respective losses which translates into higher CoT accuracy, but fails to translate into higher no-CoT accuracy.

Setup We attempt to incentivize the model to learn generalizing two-hop circuitry rather than to memorize two-hop facts directly. To that end, we split our entity triplets (e.g. (Imagine, John Lennon, Yoko Ono)) into two sets, demonstrated and undemonstrated.

1. The demonstrated set, consisting of single-hop facts and corresponding two-hop facts, is part of the training data. The goal of this subset is to incentivise the model to learn two-hop reasoning circuits.
2. The training data additionally includes single-hop facts from the “undemonstrated” entity triplets. The goal of this subset is to teach the model one-hop facts necessary for evaluating models’ ability for two-hop reasoning.
3. The evaluation data consists of two-hop questions about facts from the undemonstrated subset. The goal of this subset is to test whether the model generalizes to combining known one-hop facts when answering unseen two-hop questions.

The visualization of this dataset structure is shown in Table 1.

Results We compare the following three training data mixtures:

1. **Only atomic facts.** Training on just one-hop facts reaches perfect accuracy on one-hop questions but does not generalize to answering two-hop questions with CoT or without CoT.
2. **Atomic and two-hop no-CoT QA pairs.** Adding two-hop no-CoT QA pairs to the training dataset improves test loss on two-hop no-CoT answers compared to only training on atomic facts, but does not impact accuracy. We investigate the improvement in loss and show it is not a result of improved two-hop reasoning but is likely due to learning to conform with the no-CoT evaluation format (see Figure 1).
3. **Atomic, two-hop no-CoT and two-hop CoT QA pairs.** We additionally include CoT QA pairs in the training dataset, which further improves two-hop CoT accuracy but does not affect two-hop no-CoT accuracy. We base other interventions on this mixture and include this result in Figure 1.

The data mixture intervention fails to elicit two-hop reasoning (Figure 3). Adding two-hop QA pairs to the training dataset slightly decreases the test loss on no-CoT answers, but the test loss plateaus long before reaching zero and its decrease does not translate into accuracy improvements. We demonstrate that lower test loss compared to training only on atomic facts is not due to improved latent reasoning abilities by computing test loss on shuffled labels (see Figure 1). We believe the lower loss should be explained away as learning the no-CoT QA format used in the evaluation.

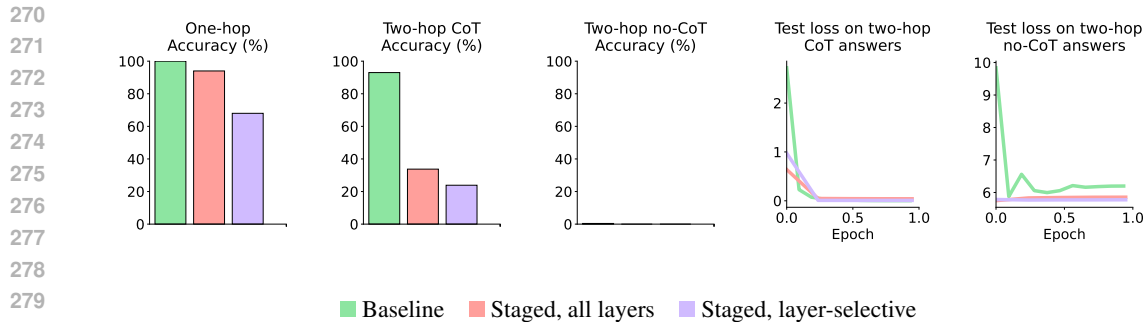


Figure 4: Performance of models trained on with different fact storage interventions across different metrics. While for the baseline, test loss on two-hop CoT answers reaches 0, loss on two-hop plateau at a much higher value. Our intervention (■ staged, layer-selective) decreases test loss slightly (rightmost plot), but but this does not translate into above-chance no-CoT accuracy (middle) and is actually *harmful* for one-hop accuracy (leftmost barplot).

5 INTERVENTION 2: FORCING FACTS TO BE STORED IN THE RIGHT ORDER

Motivation Transformers are feed-forward neural networks — a sequence of blocks that have to be traversed in a linear order for a given input. Moreover, previous work suggests that transformers store facts in a somewhat localised fashion, mostly in MLP layers of a few neighboring transformer blocks (Meng et al., 2023). Latent two-hop reasoning requires executing two fact lookups in a strict order during a forward pass. For a feed-forward neural network, this is only possible if the first fact (e.g. “the performer of Imagine is John Lennon”) is stored in an earlier block than the second fact (e.g. “the spouse of John Lennon is Yoko Ono”). Otherwise, if the first fact is stored in a later block (e.g. 20th transformer block) and the second fact in an earlier block (e.g. 10th block), by the time a model completes the first lookup to resolve the bridge entity (“John Lennon”), the forward pass can no longer use the bridge entity to look up the second fact.

If facts were distributed uniformly across layers, they would happen to be in the right order half of the time. Therefore, if layer ordering was the only reason for poor two-hop performance, one would expect two-hop accuracy to be around 50%. In practice, this should be seen as a lower bound, since some facts might be represented redundantly, more than once.

Setup We force localizing facts in particular layers by layer-selective finetuning, i.e. dividing our training distribution into three datasets and training separately on each, involving only a particular layer range at each stage:

1. *First one-hop facts* (e.g. “the performer of Imagine is John Lennon”) are learned with layers 0-12 (with other layers frozen)
2. *Second one-hop facts* (e.g. “the spouse of John Lennon is Yoko Ono”) are learned with layers 12-24 (with other layers frozen)
3. *Two-hop QA pairs* are learned with all layers updated.

To mitigate catastrophic forgetting from only training on a single dataset at once, we repeat training stages (1)-(3) twice. Moreover, our training data uses the mixture described in the previous section: training on one-hop facts and both two-hop CoT and no-CoT QA pairs.

Results We compare the following three setups:

1. **Baseline.** This is the setup from Figure 3, training on one-hop facts and both two-hop CoT and no-CoT QA pairs in a single stage with all layers trained.
2. **Staged, with all layers trained.** This setup is a sanity check to show that staged training preserves most of the baseline’s performance.
3. **Staged, layer-selective training.** This is the intervention setup.

As seen in Figure 4, forcing one-hop facts to be localized in the correct order — with the first fact stored earlier than the second one — failed to elicit two-hop reasoning. This means that correct knowledge localization in the forward pass is not enough to elicit two-hop reasoning: the model still fails to connect pieces of knowledge for answering two-hop questions.

6 INTERVENTION 3: ACTIVATION SUPERVISION FOR TWO-HOP REASONING

Motivation The cross-entropy language modeling loss, used during LLM pre-training and supervised fine-tuning, treats the LLM as a black box and only supervises how the input tokens in the prompt are mapped to output tokens. From success of CoT performance, we know that such supervision is effective in teaching models to reason in explicit CoT. Since the reasoning trace is expressed in token space, the language modeling loss provides LLMs process-based supervision, giving useful gradients for each step of reasoning. However, for reasoning in latent space, the language modeling loss only provides outcome-based feedback (whether the predicted answer is correct) and is indifferent to whether an LLM arrives at the answer via memorization or two-hop reasoning.

Setup We add an auxiliary loss \mathcal{L}_{aux} that complements outcome-based supervision from the language modeling loss with process-based feedback in the activation space. More specifically, we encourage the model to resolve the bridge entity in activation space whenever it is prompted with a two-hop question. We encourage such resolution by ensuring that a given hidden state (output of a transformer block) is either similar to a vector representation of the bridge entity or predictive of it.

We apply the auxiliary objective to the output of a single transformer block at a single token position. We sweep over several blocks to apply this loss on and choose block 10 (out of 32). To determine the token position to apply loss on, we look for the last token of the description of the bridge entity in the question, e.g. “gine” in “Who is the spouse of the singer of the song Imagine?”. Let’s call this activation vector h .

We consider two auxiliary objectives:

1. *Logit lens.* We compute logits y as $y = W_U \text{RMSNorm}(h)$, where $\text{RMSNorm}(\cdot)$ denotes the final RMSNorm (Zhang & Sennrich, 2019) layer of Llama 3 8B Instruct during training. We then compute $\mathcal{L}_{\text{aux}} = \text{CE}(e_2, y)$, where $\text{CE}(\cdot)$ is the standard cross-entropy loss and e_2 is the token corresponding to bridge entity, e.g. “John Lennon”. This is possible because we ensure all bridge entities are single-token.
2. *Embed lens.* We compute $\mathcal{L}_{\text{aux}} = -\text{CosSim}(W_E e_2, y)$, where $\text{CosSim}(\cdot)$ is the cosine similarity loss and $W_E e_2$ is the embedding of the bridge entity token.

In both cases, our final loss is computed as $\mathcal{L} = \mathcal{L}_{\text{LM}} + c\mathcal{L}_{\text{aux}}$, where \mathcal{L}_{LM} is the standard language modelling loss and the coefficient c is a hyperparameter. Based on our sweeps, we found that 0.01 and 0.1 were the best settings for logit lens and embed lens, respectively. Once again, our training data uses the setup described for Hypothesis 2 experiments: training on one-hop facts and both two-hop CoT and no-CoT QA pairs.

Results We compare the following three setups:

1. **Baseline:** This is the setup from Figure 3, training on one-hop facts and both two-hop CoT and no-CoT QA pairs with just \mathcal{L}_{LM} .
2. **Logit lens.** This is the Logit lens setup, using the best coefficient c value from a sweep.
3. **Embed lens.** This is the Embed lens setup, using the best coefficient c value from a sweep.

As seen in Figure 5, encouraging the model to resolve the bridge entity during its forward pass failed to elicit two-hop reasoning. As seen by the evaluation \mathcal{L}_{aux} , learning to resolve bridge entities during training does not generalize to resolving other bridge entities on evaluation prompts despite the training \mathcal{L}_{aux} reaching zero.

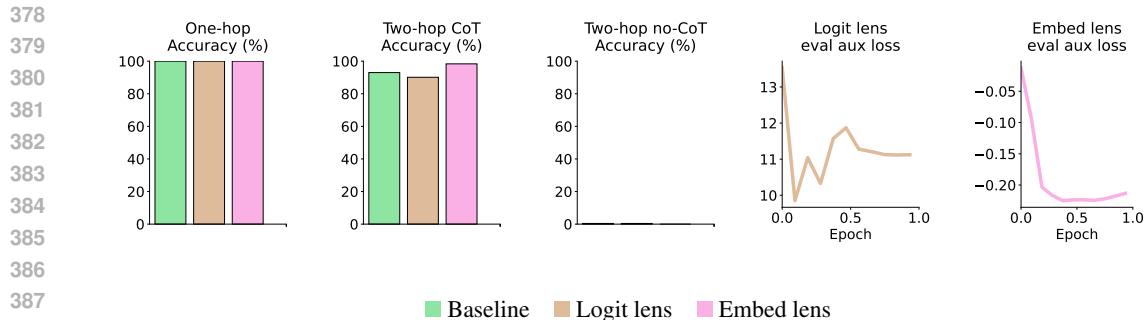


Figure 5: Performance of models trained with different auxiliary objectives across different metrics. Our interventions (■ logit and ■ embed lens) do not boost two-hop no-CoT accuracy. The two rightmost plots show empirical values of \mathcal{L}_{aux} on the test set during training for both auxiliary losses. \mathcal{L}_{aux} tends to decrease for both, but it’s either unstable (for logit lens) or tends to show signs of rapid overfitting (for embed lens). Note that a priori cross-entropy of 10 and cosine similarity of 0.2 are low values; perfect generalization would correspond to cross-entropy 0 and cosine similarity of 1.

7 LIMITATIONS

In this paper, we try to investigate the capabilities of LLMs in naturalistic settings, while controlling for confounders plaguing prior work. Reconciling the need for a clean setup and plausibility required several design choices that could be controversial.

Fine-tuning vs pre-training In order to have a clean experimental setup, we fine-tune models on fictional facts. However, one might worry that the cleanliness of this setup is fundamentally different from how knowledge is normally acquired by LLMs during pre-training. This difference might manifest in diversity of the data distribution and the scale of the training dataset.

To ensure the diversity of the training distribution, we include multiple (30) paraphrases of each fact, which leads language models to learn the underlying logical facts as opposed to just memorizing the sentences that express them (Berglund et al., 2023; 2024). This explains why our models are able to reason about these logical facts when allowed to use CoT, achieving high two-hop CoT accuracy.

Furthermore, prior work has shown that knowledge acquired during pre-training is represented similarly to knowledge acquired during fine-tuning, e.g. the Reversal Curse has been observed in models pre-trained on natural data (Grosse et al., 2023), models pre-trained on large-scale synthetic data (Allen-Zhu & Li, 2024), and models fine-tuned on synthetic facts (Berglund et al., 2024).

Ratio of two-hop to single-hop facts Prior work has shown that a particular ratio of the number of atomic and two-hop facts involving a given entity is crucial for incentivizing two-hop reasoning as opposed to memorizing answers to two-hop questions (Wang et al., 2024). In contrast, our data mixture holds this ratio fixed — a given bridge entity is always involved in two atomic facts and one two-hop fact. This might create insufficient pressure for the model to learn two-hop reasoning.

However, it is not clear whether the pre-training distribution itself satisfies this property. Future work could explore the effect of varying this ratio in naturalistic settings.

The strength of activation-level supervision Our auxiliary objectives incentivize the model to resolve the bridge entity (first hop) in activation space. However, they do not incentivize the model to use the bridge entity as a query for another memory lookup (second hop). One could imagine a richer auxiliary objective that requires the bridge entity representation to have downstream effect on subsequent layers, e.g. maximizing the gradient of the final answer w.r.t. to the representation of the bridge entity (Koh & Liang, 2017). However, such loss function would require computing second-order gradients, which is challenging to implement in distributed training setups for LLMs.

8 CONCLUSION

Previous work pointed out the existence of a *compositionality gap* — a difference in performance of LLMs at answering two-hop questions with and without CoT. In this work, we introduce a natural-language yet controlled setting for studying the compositionality gap in LLMs, where latent two-hop reasoning can be the only explanation for positive performance. We explore three groups of interventions to elicit latent two-hop reasoning: (i) a data mixture designed to incentivize learning of two-hop reasoning, (ii) forcing facts to be localized in the right order, and (iii) encouraging the bridge entity to be resolved in early layers. All of these interventions fail to improve latent reasoning ability measured by both accuracy and loss, while achieving near-perfect two-hop CoT accuracy. At the very least, we show that eliciting latent two-hop reasoning in LLMs is not trivial: we believe our experiments tried picking the lowest-hanging fruit and found that it is all sour.

Further, our results lead us to believe that previous work might have significantly overestimated the extent to which latent two-hop reasoning occurs in LLMs. While it is undeniable that latent two-hop reasoning is representable by transformers (Wang et al., 2024), we conjecture that current LLMs are unlikely to actually perform latent two-hop reasoning. If LLMs did perform two-hop reasoning, they would have more than chance-level loss on answers to two-hop questions that they can answer with near-perfect accuracy using explicit CoT. In line with past work on fundamental limitations of LLMs (Berglund et al., 2024), we call this failure of LLM reasoning the Two-Hop Curse.

REFERENCES

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation, 2024. URL <https://arxiv.org/abs/2309.14402>.
- Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=oVTkOs8Pka>. Survey Certification, Expert Certification.
- Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms, 2023. URL <https://arxiv.org/abs/2309.00667>.
- Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on “a is b” fail to learn “b is a”. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=GPKTIktA0k>.
- Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries, 2024. URL <https://arxiv.org/abs/2406.12775>.
- Joe Carlsmith. Scheming ais: Will ais fake alignment during training in order to get power?, 2023. URL <https://arxiv.org/abs/2311.08379>.
- Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into ai agents, 2024. URL <https://arxiv.org/abs/2401.13138>.
- Xin Chen, Hanxian Huang, Yanjun Gao, Yi Wang, Jishen Zhao, and Ke Ding. Learning to maximize mutual information for chain-of-thought distillation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp.

- 486 6857–6868, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational
 487 Linguistics. doi: 10.18653/v1/2024.findings-acl.409. URL [https://aclanthology.org/
 488 2024.findings-acl.409](https://aclanthology.org/2024.findings-acl.409).
- 489 Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to inter-
 490 nalize cot step by step, 2024. URL <https://arxiv.org/abs/2405.14838>.
- 491
- 492 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
 493 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony
 494 Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark,
 495 Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere,
 496 Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris
 497 Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong,
 498 Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny
 499 Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino,
 500 Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael
 501 Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Ander-
 502 son, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah
 503 Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan
 504 Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Ma-
 505 hadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy
 506 Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak,
 507 Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Al-
 508 wala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini,
 509 Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der
 510 Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo,
 511 Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Man-
 512 nat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova,
 513 Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal,
 514 Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur
 515 Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhar-
 516 gava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong,
 517 Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic,
 518 Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sum-
 519 baly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa,
 520 Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang,
 521 Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende,
 522 Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney
 523 Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom,
 524 Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta,
 525 Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petro-
 526 vic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang,
 527 Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur,
 528 Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre
 529 Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha
 530 Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay
 531 Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda
 532 Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew
 533 Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita
 534 Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh
 535 Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De
 536 Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Bran-
 537 don Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina
 538 Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai,
 539 Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li,
 540 Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana
 541 Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil,
 542 Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Ar-
 543 caute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco
 544 Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella

- 540 Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory
541 Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang,
542 Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Gold-
543 man, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman,
544 James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer
545 Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe
546 Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie
547 Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun
548 Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal
549 Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva,
550 Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian
551 Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson,
552 Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Ke-
553 neally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel
554 Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mo-
555 hammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navy-
556 ata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong,
557 Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli,
558 Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux,
559 Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao,
560 Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li,
561 Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott,
562 Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Sa-
563 tadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lind-
564 say, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang
565 Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen
566 Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho,
567 Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser,
568 Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Tim-
569 othy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan,
570 Vinay Satish Kumar, Vishal Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu
571 Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouwaziz, Will Con-
572 stable, Xiaocheng Tang, Xiaofang Wang, Xiaojuan Wu, Xiaolan Wang, Xide Xia, Xilun Wu,
573 Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi,
574 Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef
575 Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.
576 URL <https://arxiv.org/abs/2407.21783>.
- 575 Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter
576 West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xi-
577 ang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: limits of transformers
578 on compositionality. In *Proceedings of the 37th International Conference on Neural Information
579 Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.
- 580 Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing
581 the mystery behind chain of thought: A theoretical perspective. In *Thirty-seventh Conference on
582 Neural Information Processing Systems*, 2023. URL [https://openreview.net/forum?
583 id=qHrADgAdYu](https://openreview.net/forum?id=qHrADgAdYu).
- 584 Roger Grosse, Juhun Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit
585 Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilè Lukošiušė, Karina Nguyen,
586 Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. Studying large lan-
587 guage model generalization with influence functions, 2023. URL [https://arxiv.org/
588 abs/2308.03296](https://arxiv.org/abs/2308.03296).
- 589 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn
590 Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math
591 dataset. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information
592 Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL [https://
593 datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/
2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf](https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf).

- 594 Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner,
595 Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger
596 language models with less training data and smaller model sizes. In Anna Rogers, Jordan Boyd-
597 Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics:
598 ACL 2023*, pp. 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL [https://aclanthology.org/2023.
600 findings-acl.507](https://aclanthology.org/2023.findings-acl.507).
- 601 Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
602 and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- 603 Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from
604 learned optimization in advanced machine learning systems, 2021. URL [https://arxiv.
605 org/abs/1906.01820](https://arxiv.org/abs/1906.01820).
- 606 Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In
607 *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML’17*,
608 pp. 1885–1894. JMLR.org, 2017.
- 609 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large
610 language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on
611 Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA, 2024. Curran Associates
612 Inc. ISBN 9781713871088.
- 613 Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Her-
614 nandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuėtė, Karina
615 Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson,
616 Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Tim-
617 othy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan
618 Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought rea-
619 soning, 2023. URL <https://arxiv.org/abs/2307.13702>.
- 620 Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual
621 associations in gpt, 2023. URL <https://arxiv.org/abs/2202.05262>.
- 622 William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision trans-
623 formers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023a. doi:
624 10.1162/tacl.a.00562. URL <https://aclanthology.org/2023.tacl-1.31>.
- 625 William Merrill and Ashish Sabharwal. A logic for expressing log-precision transformers. In
626 *Thirty-seventh Conference on Neural Information Processing Systems, 2023b*. URL <https://openreview.net/forum?id=uR8TtWCIsr>.
- 627 William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought,
628 2024. URL <https://arxiv.org/abs/2310.07923>.
- 629 Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures
630 for grokking via mechanistic interpretability, 2023. URL [https://arxiv.org/abs/2301.
631 05217](https://arxiv.org/abs/2301.05217).
- 632 Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning
633 perspective, 2024. URL <https://arxiv.org/abs/2209.00626>.
- 634 OpenAI. Openai o1 system card. System card, OpenAI, Sept 2024.
- 635 Jacob Pfau, William Merrill, and Samuel R. Bowman. Let’s think dot by dot: Hidden computation
636 in transformer language models, 2024. URL <https://arxiv.org/abs/2404.15758>.
- 637 Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Mea-
638 suring and narrowing the compositionality gap in language models. In Houda Bouamor, Juan
639 Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP
640 2023*, pp. 5687–5711, Singapore, December 2023. Association for Computational Linguistics.
641 doi: 10.18653/v1/2023.findings-emnlp.378. URL [https://aclanthology.org/2023.
642 findings-emnlp.378](https://aclanthology.org/2023.findings-emnlp.378).

- 648 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien
649 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a
650 benchmark. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=Ti67584b98>.
651
- 652 Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the
653 few-shot paradigm, 2021. URL <https://arxiv.org/abs/2102.07350>.
654
- 655 Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann
656 Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-
657 thought helps mainly on math and symbolic reasoning, 2024. URL <https://arxiv.org/abs/2409.12183>.
658
- 659 Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. What formal languages
660 can transformers express? a survey. *Transactions of the Association for Computational Linguistics*,
661 12:543–561, 2024. doi: 10.1162/tacl.a.00663. URL <https://aclanthology.org/2024.tacl-1.30>.
662
- 663 Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don’t always
664 say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the*
665 *37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook,
666 NY, USA, 2024. Curran Associates Inc.
667
- 668 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,
669 Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von
670 Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Ad-*
671 *vances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.,
672 2017. URL [https://proceedings.neurips.cc/paper_files/paper/2017/](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf)
673 [file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf).
- 674 Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked transformers are implicit reasoners: A
675 mechanistic journey to the edge of generalization, 2024. URL <https://arxiv.org/abs/2405.15071>.
676
- 677 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc
678 Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models,
679 2023. URL <https://arxiv.org/abs/2201.11903>.
680
- 681 Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large lan-
682 guage models latently perform multi-hop reasoning? In Lun-Wei Ku, Andre Martins, and
683 Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Com-*
684 *putational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229, Bangkok, Thailand, August
685 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.550. URL
686 <https://aclanthology.org/2024.acl-long.550>.
- 687 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov,
688 and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question
689 answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun’ichi Tsujii (eds.), *Proceed-*
690 *ings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–
691 2380, Brussels, Belgium, October–November 2018. Association for Computational Linguistics.
692 doi: 10.18653/v1/D18-1259. URL <https://aclanthology.org/D18-1259>.
- 693 Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of language models: Part 2.1,
694 grade-school math and the hidden reasoning process, 2024. URL <https://arxiv.org/abs/2407.20311>.
695
- 696 Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1, 2024. URL
697 <https://arxiv.org/abs/2407.06023>.
698
- 699 Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with
700 reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Ad-*
701 *vances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/](https://openreview.net/forum?id=_3ELRdg2sgI)
[forum?id=_3ELRdg2sgI](https://openreview.net/forum?id=_3ELRdg2sgI).

702 Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Good-
703 man. Quiet-STAR: Language models can teach themselves to think before speaking. In *First*
704 *Conference on Language Modeling*, 2024. URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=oRXPiSOGH9)
705 [oRXPiSOGH9](https://openreview.net/forum?id=oRXPiSOGH9).
706
707 Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019. URL [https://](https://arxiv.org/abs/1910.07467)
708 arxiv.org/abs/1910.07467.
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755