How Fine-Tuning Data Influences Reliance on Irrelevant Context in Math Reasoning

Anonymous authors
Paper under double-blind review

Abstract

Large language models have been shown to rely on and get distracted by irrelevant context (IC) when solving mathematical reasoning tasks. While there exist many works that show this failure mode, it is not well understood what factors contribute to such reliance on irrelevant context. In this work, we take a first step towards understanding the origins of this failure mode. First, we perform ablations on the irrelevant context and show that models often learn shortcut correlations between artifacts and operations without understanding the requirement of the question, which then causes them to rely on irrelevant context. Next, we discover and show that long chain-of-thought supervision exacerbates reliance on irrelevant context. We also show that introducing irrelevant context during fine-tuning can make these models more robust against irrelevant context. Interestingly, this is true even if the added IC is fully out-of-domain with respect to the original fine-tuning set. We conduct comprehensive generalization tests and also evaluate on more realistic and challenging ICs. Finally, we show that inclusion of such IC within the fine-tuning set does not negatively impact performance on clean test sets, while improving robustness by overcoming reliance on ICs.

1 Introduction

Deep neural networks have shown impressive performance on many tasks, including those that require complex reasoning ability (Cobbe et al., 2021; Hendrycks et al., 2021). However, these models have been shown to be sensitive to minor perturbations in the input provided. Shi et al. (2023) first showed that if these models were provided grade school math reasoning questions with irrelevant context unrelated to the question, these models would produce incorrect answers on questions they would get right without the irrelevant context present in the questions. While all prior work have shown this phenomenon (Shi et al., 2023; Li et al., 2024; Yang et al., 2025), it is not understood what factors contribute to such reliance on irrelevant context.

In this work, we answer why these models rely on irrelevant context by performing a more in-depth analysis of the training data used to finetune these models. We show that models learn correlations between entities and arithmetic operations without scope resolution, making them susceptible to reliance on irrelevant context. We also present insights regarding the nature of training samples and show that training samples containing long chain-of-thought (CoT) reasoning steps encourage a model to rely on irrelevant information when making decisions or predictions. We also show that increasing test-time thinking tokens (CoT tokens) (Muennighoff et al., 2025) increases IC reliance. Next, we show that one can simply overcome such reliance on irrelevant context by including out of domain and irrelevant information within questions of the fine-tuning set. We also test the efficacy, efficiency, and generalizability of including such distractors within the fine-tuning set across different settings and model families. We list our primary contributions below:

- We show that current LLMs simply learn shortcut correlations between entities observed in the input and operations in the CoT supervision signals. This causes the network to rely on irrelevant context.
- We discover that long CoT supervision signals contribute strongly to irrelevant context reliance, encouraging the model to generate verbose CoTs during test time and making it more likely to draw in irrelevant information into the reasoning steps. We show that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: An example of `Extended GSM-IC`: irrelevant context added to a problem of GSM8k.

Question from GSM8k Test Set: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?
Question with Irrelevant Context: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. Janet spends 50 hours sleeping per week. How much in dollars does she make every day at the farmers' market?

simply scaling up test-time thinking tokens (CoT tokens) (Muennighoff et al., 2025) also exacerbates IC reliance.

- We show that inclusion of irrelevant context in the fine-tuning data, even if out-of-domain, can improve robustness to in-domain irrelevant context. We also perform a generalization assessment and show that such techniques generalize well to other unseen distractors–in terms of distractor format or domain relevance. We also assess these models on more challenging and realistic ICs.

We note that we restrict our scope to evaluating fine-tuned models instead of black-box models whose pre-training data mixture is unknown. This way, one can draw conclusions about how training data impacts reliance on irrelevant context.

2 Background and Setup

Reasoning. Studying the ability of large language models to perform reasoning has been an area of strong importance and interest. Models capable of reasoning can solve more complex tasks that require multiple steps, improving the overall ability and reliability of these networks. Recent work has shown that these models are capable of solving reasoning tasks in different domains (Cobbe et al., 2021; Wei et al., 2022a). Since then, the reasoning ability of these models has improved vastly, mostly stemming from improvements in pre-training (Brown et al., 2020), post-training, and test-time scaling techniques (Brown et al., 2024; Snell et al., 2025; Muennighoff et al., 2025).

Chain-of-Thought (CoT). To enhance the reasoning ability of large language models, they are trained or prompted to produce a chain of reasoning steps before arriving at an answer. These steps are often termed as chain-of-thought (Cobbe et al., 2021; Nye et al., 2021; Wei et al., 2022b).

Irrelevant Context. Recent work has shown that large language models, even those at the frontier, often rely on irrelevant context when solving these tasks (Shi et al., 2023; Li et al., 2024; Mirzadeh et al., 2025; Yang et al., 2025). While prior works show that this is an existing failure mode, it is unclear as to why this happens.

Spurious Correlations. Spurious Correlations are correlations that the model learns between spurious, causally irrelevant features and ground truth labels (Arjovsky et al., 2019; Sagawa et al., 2020; Kirichenko et al., 2023). The learning of such spurious correlations during training is problematic because such correlations may no longer hold at test time or the spurious feature may become correlated with a different label altogether, causing the network to malfunction when deployed in the real world.

Evaluation Suites used and their generation and characteristics. In this work, we use many different evaluation sets based on the mathematical reasoning task we are concerned with. While we use existing benchmarks with irrelevant context, we also create our own based on templates available. We provide details regarding their generation below for easy reproducibility.

- **GSM8K** is a widely used benchmark for training and evaluating large language models on grade school math reasoning problems. One popular benchmark for evaluating models on GSM8k but with irrelevant context is the GSM-IC benchmark (Shi et al., 2023). In the GSM-IC benchmark, different IC sentences are injected into the standard GSM8k train set. However, this benchmark is based on only 100 samples, making it unreliable for robustly drawing insights. Thus, building on it, we create the following three evaluation suites:

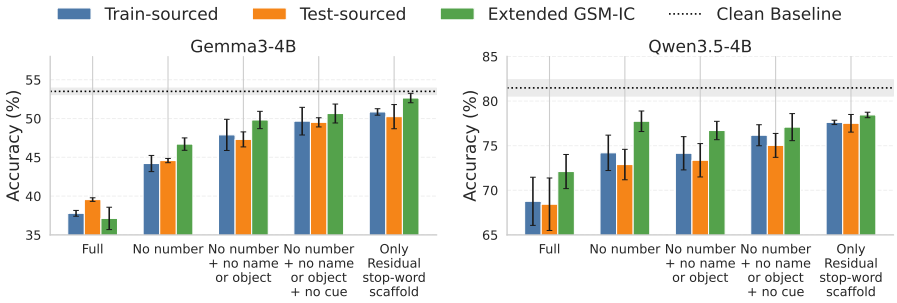


Figure 2: Irrelevant numbers and cues are mapped to operations without resolving scope, hurting performance. In this figure, baseline represents performance of the same model but on the standard test set of GSM8k, without any IC.

  - `Extended GSM-IC`: We utilize the GSM-IC template and extend the benchmark to cover all 1319 test samples of the GSM8k dataset. We present an example in Table 1.
  - `Train-Sourced Distractor Set`: To the GSM8k test set, we introduce distractors sourced from the train set.
  - `Test-Sourced Distractor Set`: To the GSM8k test set, we introduce distractors sourced from the test set, but ensure that they are not sourced from the same questions.
- **MATH500.** We use MATH500 benchmark (Lightman et al., 2024), derived from MATH benchmark Hendrycks et al. (2021). We introduce irrelevant context to assess for IC reliance.

In this work, we study open-weight models (Qwen Team, 2026; Team et al., 2025) fine-tuned on the GSM8k and MATH datasets. We provide additional details, including training details, in the Appendix. To compute the degree of reliance on irrelevant context, we compare the accuracies attained on these benchmarks with irrelevant context against the accuracy attained on the standard (clean) GSM8k test set. We also compute IC uptake values that measure the number of test samples where numeric tokens within the IC was used within the model's CoT. We note that GSM8k is the only available mathematical reasoning benchmark with irrelevant context variants. To ensure the generality of these results, we create a variant of the MATH500 dataset that contains IC. Finally, we note that we use the terms irrelevant context (IC) and distractors interchangeably throughout the text.

3 Cause of Irrelevant Context Reliance

3.1 Shortcut Learning: Entity-to-Operation Correlations

One plausible reason behind the reliance of these models on irrelevant context is that these models may simply be learning observed correlations between entities in the input prompt and operations in the Chain-of-Thought (CoT) within the training set during fine-tuning, without considering the scope or requirement defined by the question. So when it encounters such an entity during test, it simply performs a corresponding irrelevant operation instead of robustly reasoning and considering which operations need to be performed based on the question. To examine this, we test a model fine-tuned on the standard GSM8k train set on two evaluation sets:

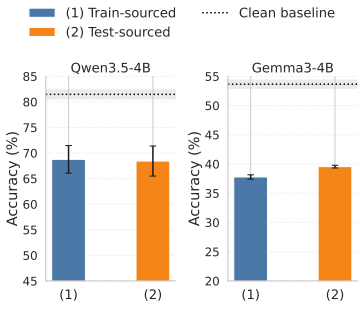


Figure 1: Train and Test set sourced distractors have a similar impact on testing performance.

1. `Train-sourced`: We create a GSM-IC style distractor set with samples from the GSM8k test set where we source irrelevant sentences from the **train** set.

2. `Test-sourced`: We create a GSM-IC style distractor set with samples from the GSM8k test set where we source irrelevant sentences from the **test** set.

On both sets, the model obtains similar testing performances, indicating that the model is not learning correlations between train-time observed whole sentences/phrases and operations (Fig. 1).

Field	Content
Gold answer	2
Question with IC	4 adults and 8 children are to share 8 packets of chocolate bars. Each packet contains 5 chocolate bars. Kim's mother bought 10 more T-shirts. If each adult gets 6 chocolate bars and the rest are to be shared equally among the children, how many will each child get?

Resultant Chain-of-Thought	`There are a total of 4 x 6 = «4*6=24»24 chocolate bars for the adults.`
	`So, there are 8 x 5 + 10 = «8*5+10=50»50 chocolate bars in total.`
	`Therefore, there are 50 - 24 = «50-24=26»26 chocolate bars for the children.`
	`Hence, each child will get 26/8 = «26/8=3.25»3.25 chocolate bars.`
	`The final answer is: 3.25` (Incorrect)

Question with IC but number removed	4 adults and 8 children are to share 8 packets of chocolate bars. Each packet contains 5 chocolate bars. Kim's mother bought more T-shirts. If each adult gets 6 chocolate bars and the rest are to be shared equally among the children, how many will each child get?

Resultant Chain-of-Thought	`There are a total of 4 x 6 = «4*6=24»24 chocolate bars for the adults.`
	`So, there are 8 x 5 - 24 = «8*5-24=16»16 chocolate bars for the children.`
	`Therefore, each child will get 16/8 = «16/8=2»2 chocolate bars.`
	`The final answer is: 2` (Correct)

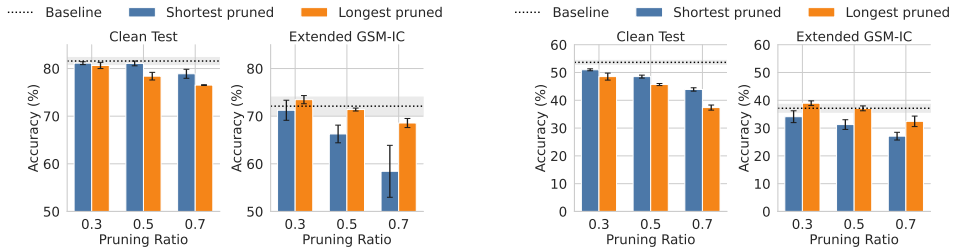
Table 2: Simply removing the number from the irrelevant context overcomes reliance on IC, implying that the model learns shortcut correlations between entities and operations. Response Generated from Qwen3.5-4B fine-tuned on standard GSM8k train set.

Finer-Grained Correlations

While the model is not mapping entire sentences to operations, we suspect that it might be learning more fine-grained correlations between entities and operations. To test this, we perform ablations on the irrelevant context sentences, where we remove certain components from the irrelevant context within the evaluation set. By systematically removing components from the irrelevant sentence, it will become clearer as to which components within the irrelevant context are being mapped to operations during inference without considering the overall scope or requirement of the question. To do so we remove the following components from the irrelevant context systematically:

- First we remove any ***numbers*** from the irrelevant sentence.
- Next, we remove any ***names or objects*** from the irrelevant sentence.
- Next, we remove ***operation cues*** such as "`half`" or "`remaining`" from the sentence. We present a complete list of the operation cues in the appendix (Sec. C).
- Finally, we remove any words that are not stop-words. This step comprises removing ***verbs, adjectives and adverbs*** such as "`rowing`" or "`current`".

In Fig. 2 and Table 2, we observe that simply removing the numbers from the irrelevant sentence yields notable improvements in performance, up to 10% points (by comparing `Full` and `No Numbers`). Note that simply removing the number will cause almost no impact on the question length. We also observe that as we continue to remove these components from the irrelevant sentence while maintaining the remainder of the irrelevant sentences, performance improves notably matching the performance of the model on the clean evaluation set. Such an observation is interesting because it shows that even *irrelevant objects or names*, without being in conjunction with a number, can cause the network to perform incorrect or additional arithmetic operations. These observations thereby imply that fine-tuned models map artifacts to operations during training, without considering the



(a) Qwen3.5-4B on GSM8k (b) Gemma3-4B on GSM8k

Figure 3: Pruning samples with short CoT significantly hurts performance on Extended GSM-IC. Prune ratio indicates fraction of the standard GSM8k train set that has been pruned. Baseline indicates performance without any pruning.

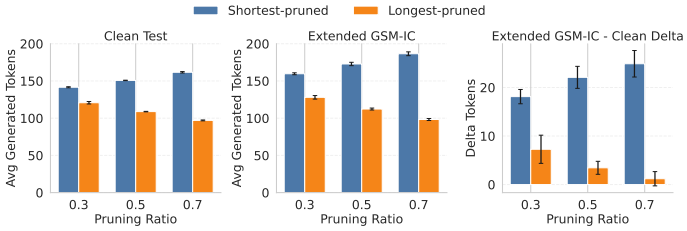


Figure 4: Test time impact of two different data pruning strategies in train time. Long CoT Supervision (`Shortest Pruned`) teaches the model to write verbose reasoning steps during test time, encouraging it to draw in irrelevant information. Results shown with Gemma3-4B. Qwen3.5-4B results shown in the Appendix (Fig. 11).

scope or requirement of the question. Such shortcut learning causes these models to rely on irrelevant context during testing.

3.2 Sources of Shortcut Learning during Fine-tuning

In this section, we aim to understand what causes such shortcut learning that encourages reliance on irrelevant context. More specifically, we aim to identify if there are certain parts or characteristics of the fine-tuning data that contribute disproportionately to reliance on irrelevant context and if one can simply remove them. Empirically, it is evident that models rely on irrelevant context if they simply perform operations based on entities and patterns seen during training time. Through our experiments, we identify the following characteristics of the fine-tuning data that biases a network to rely on irrelevant context.

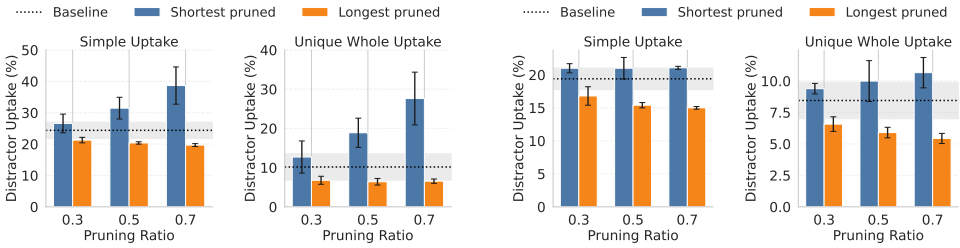
3.2.1 *Long* Chain-of-Thought Supervision

We discover that **language models supervised to produce more verbose CoTs are more likely to rely on irrelevant context**. To show this, we train two sets of models, trained on increasingly pruned versions of the standard GSM8k train set:

- `Longest pruned`: We prune samples with the longest CoTs first as we move towards samples with the shortest CoTs as we prune more (Short CoT supervision).
- `Shortest pruned`: We prune samples with the shortest CoTs first as we move towards samples with the longest CoTs as we prune more (Long CoT supervision).

We then compare their performances on the Extended GSM-IC test set and also the standard (Clean) GSM8k test set. We present the results in Fig. 3.

We observe that `Shortest Pruned` (thereby maintaining samples containing long CoTs) significantly hurts performance on the Extended GSM-IC test set (blue bars in Fig. 3 (a) Right and (b) Right). In contrast, `Longest Pruned` (thereby maintaining samples with short CoTs) achieves comparable accuracy to the unpruned baseline (orange bars in Fig. 3 (a) Right and (b) Right). Interestingly, we observe the **opposite** effect on the standard, clean GSM8k test set. `Longest Pruned`



(a) Qwen3.5-4B on Extended GSM-IC (b) Qwen3.5-4B on MATH500-IC

Figure 5: Long CoT Supervision (`Shortest Pruned`) increases uptake of irrelevant context.

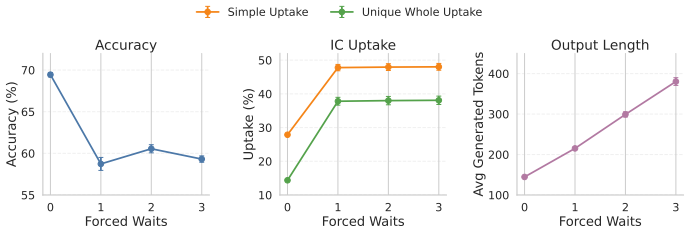


Figure 6: Scaling up test time thinking tokens (CoT) increases IC uptake and can hurt overall performance. Results shown with Qwen2.5-7B on Extended GSM-IC.

hurts performance more (orange bars in Fig 3 (a) Left and (b) Left), implying that training samples with long CoTs encourage the model to learn math-relevant skills, but by looking at the blue bars in Fig 3 (a) Right and (b) Right, we can infer that long CoTs also exacerbate irrelevant context reliance. These observations collectively confirm that **long CoTs contribute both to general performance and reliance on ICs.**

We also observe that long CoT supervision during training (`Shortest Pruned`) teaches the model to produce longer CoTs during test time, thereby encouraging the model to draw in more information from the input prompt, which may contain irrelevant components (blue bars in Fig. 4). Such an observation explains why maintaining samples with long CoTs significantly hurts performance on the Extended GSM-IC test set (blue bars in Fig. 3 (a) Right and (b) Right). On the other hand, short CoT supervision (`Longest Pruned`) encourages the model to produce shorter CoTs during test time, restricting the amount of tokens used to come to the answer (orange bars in Fig. 4). This discourages the model from drawing in irrelevant information from the input.

We note that this is the first work to show that long CoT supervision and long CoT generation contribute to irrelevant context reliance. Prior work has shown that increasing CoT tokens can significantly improve performance (Muennighoff et al., 2025) on clean test sets, which we also observe in Fig. 3 (a) Left and (b) Left. However, once irrelevant context comes into the picture, such an increase in CoT tokens can be harmful (Fig. 3 (a) Right and (b) Right).

We repeat this experiment with the MATH500 benchmark (Lightman et al., 2024; Hendrycks et al., 2021) that we discuss in Sec. 2. We observe similar patterns, where long-CoT pruning increases uptake of irrelevant context, as shown in Fig. 5. We also show the uptake values from our original GSM8k setup. We measure simple uptake by checking whether the exact numeric tokens in the distractor appear within the generate response. We measure unique uptake the same way but exclude those samples whose question has the same numeric tokens as the distractor sentence.

Simple Test-Time Scaling. Recent work by Muennighoff et al. (2025) shows that simply increasing the number of test time thinking tokens can improve the performance on math reasoning tasks notably. They do this by simply appending "wait" tokens as the model appears to complete its chain-of-thought, thereby extending its thinking process. We perform a similar experiment with Qwen2.5-7B fine-tuned on GSM8k and append "wait" tokens before the model outputs the final answer. We observe that such scaling up of test time tokens can increase uptake of IC and even hurt performance when the test set contains irrelevant context, as shown in Fig. 6. This further validates our finding that longer CoTs can exacerbate IC reliance.

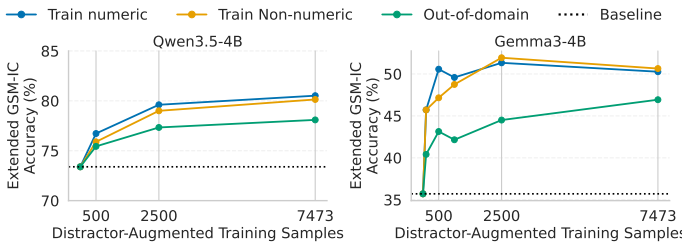


Figure 7: Performance on Extended GSM-IC scales well as we increase the number of fine-tuning samples that contain irrelevant context in them. Standard GSM8k train set contains 7,473 samples. In the plot above, we only vary the number of samples containing IC - all models are trained on all 7,473 samples (with or without IC).

Sample Difficulty is Not a Source. One might speculate that more difficult samples may tend to have longer CoTs, and thus, sample difficulty can be a contributing factor to irrelevant context reliance. We show that sample difficulty does **not** contribute to irrelevant context reliance in Fig. 9 and Sec. D in the appendix, where we show that pruning by sample difficulty does not have the same impact as pruning by CoT length, and also produces no notable patterns. We discuss this in greater detail in the appendix. This further reinforces our claim that long CoT supervision (and not sample difficulty), where the model is trained to output longer CoT reasoning steps, exacerbates irrelevant context reliance.

3.2.2 *Lack* OF APPROPRIATE SUPERVISION IN FINE-TUNING DATA

From our initial analysis, it is evident that certain parts of the training distribution (Long CoT Supervision) contribute heavily to irrelevant context reliance. However, we also show that simply removing these samples does not yield notable improvements in performance. This is because these long-CoT samples also contain training signals useful to good reasoning ability. In other words, **Training samples that exacerbate IC reliance also improve test time accuracy on clean test sets.** Thus, we look to modify the training distribution by adding supervision signals instead of removing them. In other words, we aim to add more information to the training data such that the model learns the ability to overcome irrelevant context. We discuss this in more detail in the subsequent sections.

4 CREATING ROBUST FINE-TUNING DISTRIBUTIONS

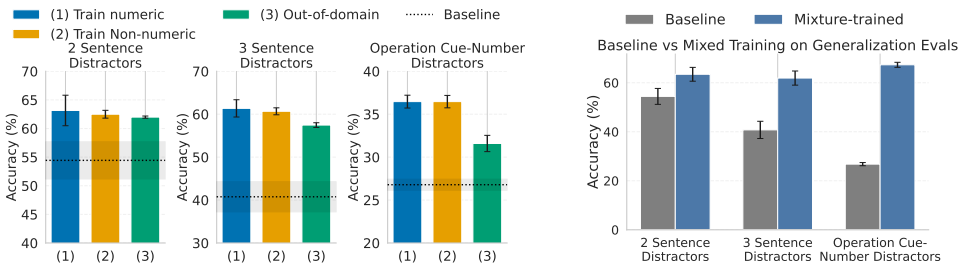
Prior work has shown that once samples with irrelevant context are introduced into the training data, these models become robust to irrelevant context reliance (Yang et al., 2025). However, they show this in a synthetic setting derived from Ye et al. (2025) and do not show the generalizability of these distractors to different kinds of distractors in terms of domain or form. In other words, distractors are drawn from the same synthetic generator, are within the same domain, and take the same form - a sentence. In this work, we perform a more granular study on the standard GSM8k dataset. We also consider more realistic and challenging irrelevant contexts, conduct more robust generalizability tests, and fine-tune with mixtures of distractors of different forms.

4.1 WHAT CONSTITUTES AN EFFECTIVE TRAINING SAMPLE FOR ROBUSTNESS TO IRRELEVANT CONTEXT?

We aim to better understand what characteristics of the irrelevant context, when introduced into the fine-tuning set samples, enable models to become more robust. To do so, we conduct the following tests by altering the nature of the distractors added to the fine-tuning dataset. Below, we discuss the three different types of distractors we study:

(i) `Train numeric`: The irrelevant contexts have *numbers* in them and are sourced from the *training* set.

(ii) `Train Non-numeric`: The irrelevant contexts have *no numbers* in them but are in-domain in that they are drawn from the training set.



(a) (b)

Figure 8: (a) Single sentence distractors introduced during fine-tuning generalize to multi-sentence distractors and improve performance over baseline notably. In this figure, the baseline model is fine-tuned on the standard GSM8k train set, without any distractors/ICs introduced into the fine-tuning set. Results shown with Qwen3.5-4B. (b) Training with a mixture of distractors creates models that are robust to all types of distractors. Results shown with Qwen3.5-4B.

(iii) `Out-of-domain`: The irrelevant contexts are drawn from a different domain and are *not relevant* to the task *at all*. Example IC sentences: *"The breeze carried the scent of pine and soil."*, *"The boat rocked gently beside the dock."*.

In all three categories, the distractor takes the form of one additional irrelevant sentence introduced into the samples of the standard GSM8k fine-tuning set. We also vary the number of samples within the GSM8k train set that contain the ICs/distractors. Finally, we evaluate on the Extended GSM-IC dataset. We observe that irrelevant contexts that are in-domain provide the same performance, regardless of whether they contain numbers or not (Fig. 7). On the other hand, if the irrelevant contexts are drawn from a different domain altogether (iii), they can also significantly improve robustness to irrelevant context (up to 12% points, although not as great as performance when the irrelevant contexts are in-domain (i & ii)). Such findings confirm the following insight:

> *To overcome irrelevant context reliance, one does not have to train with in-domain distractors alone but can use* **any** *irrelevant context sourced from anywhere, even if* **Out-of-domain**. *Such a finding provides an effective and easy way to improve the general robustness of a model to irrelevant context, regardless of the setting or dataset on which it is tested or fine-tuned.*

Impact of Sample Count. We observe in Fig. 7 that in our two settings, including distractors within the fine-tuning set is highly sample efficient, with just 500 samples (6.6% of all samples) including irrelevant context improving performance notably. This is true for distractors of all natures - even those that are sourced from different domains altogether.

4.2 Generalization to other ICs

In Section 4.1, we show that including an irrelevant sentence per sample in the fine-tuning dataset, sourced from different distributions/domains, makes the model more robust to irrelevant context during test time. Of particular interest was the case of out-of-domain distractors included within the fine-tuning set that transferred well to mathematical distractors (`Out-of-domain`) in Fig. 7. In this section, we further assess the transferability and generalizability of distractors included during training time. More specifically, we try to understand: When fine-tuning with distractor sentences, does the ability to overcome irrelevant context transfer to other distractors? In other words, we try to answer the question: When fine-tuning with single sentence-level irrelevant context, does the network learn the robust ability to overcome irrelevant context, or does it learn a shortcut that tells the model to ignore a sentence and then proceed with the chain-of-thought reasoning steps? To test this robustly, we evaluate on the following three settings. For simplicity, we evaluate models fine-tuned with all training samples containing IC in Fig. 7.

S1: Multiple scattered distractor sentences. In this test set, there are multiple distractor/irrelevant sentences in the question instead of just one irrelevant sentence (which was the distribution the model was fine-tuned on). We also scatter these sentences across the question wherever possible, instead of grouping them together. Interestingly, simply fine-tuning with single sentence-level distractors

teaches the model robust reasoning ability such that it overcomes irrelevant context of different nature - multiple sentences split across different parts of the input to the model (Fig. 8a (Left and Middle)).

S2: Non-sentence-level distractors: numbers and operation cues. Next, we extend our analysis to more realistic settings where the irrelevant context takes the form of random numbers and operation cue words scattered across the question instead of full distractor sentences. Such a setting is far more adversarial but also more realistic - a more likely distractor in an input is an additional (irrelevant) word or number that the user might enter by mistake as an input prompt. Interestingly, fine-tuning with a single sentence-level distractor still generalizes to such distractors, although the benefits are weaker (Fig. 8a (Right)).

S3: Impact on standard GSM8k test performance. We observe that including irrelevant context (even if out-of-domain) within the question during fine-tuning does not impact performance on the standard GSM8k test set that does not include explicit distractors/irrelevant context. For our Qwen3.5-4B setup, fine-tuning with `Train Numeric`, `Train Non-Numeric`, and `Out-of-domain` gives us accuracies of 81.60, 81.24, and 81.57, respectively on the standard GSM8k test set. Note that Qwen3.5-4B trained on standard GSM8K without any distractors attains an accuracy of 81.47 on the GSM8k test set.

4.3 Training with Mixture of Distractors

In Fig. 8a, we observe that training with one sentence distractors can generalize to multiple sentences really well. However, while there are performance improvements on distractors comprising scattered numbers and operation cues, the improvements over the baseline are not significant. This begs the question: *Can one train with multiple distractor types introduced into the fine-tuning set? Will such fine-tuning generalize to all kinds of distractors?* To test this, we introduce a mixture of distractors into the fine-tuning set where in 50% of the samples, we introduce a sentence distractor and in the other 50%, we introduce scattered irrelevant number and operation cue distractors. We observe that training with a mixture of distractors generalizes to all kinds of distractors included in the training set, improving the model robustness generally, as shown in Fig. 8b.

5 Conclusion

Summary. We identify that large language models rely on irrelevant context due to shortcuts they learn between train-time entities and operations without scope resolution. We also show that *long* CoT supervision exacerbates reliance on irrelevant context. Finally, we show that training with IC, even if fully out-of-domain, can overcome reliance on irrelevant context. We conduct generalizability tests of such a robustness measure and also test these models on more challenging ICs.

Applying Insights to General Practice. This work holds several novel insights that can aid in better training of robust (reasoning) models. Below, we discuss how these can be used in practice:

- **Injection of Irrelevant Artifacts within the fine-tuning data.** We show that injection of irrelevant artifacts, even when sourced from completely different domains, can improve robustness of these models. This simple technique does not impact performance on standard, clean evaluation sets, making it a simple and crucial step for fine-tuning reasoning models, analogous to augmentation/transformation based regularization in image classification tasks.
- **Practices for Data Efficient Fine-Tuning.** Data efficient practices for compute-constrained settings often remove samples that are redundant and do not impact test-time performance. In Fig. 3, we show that pruning short CoT train samples does not impact clean test performance much, which might give a practitioner the illusion of nothing being lost. Through our observations, however, we show that short CoT train samples are critical for robustness to irrelevant context and thus must be maintained.
- **Practices for (Synthetic) Data Generation/Creation.** We have shown that long CoT samples often encourage shortcut learning. Such an insight can aid practitioners in being more careful when creating or generating fine-tuning data for their task, by injecting irrelevant information within the fine-tuning data in case the fine-tuning samples have long CoTs.

AI USE STATEMENT

AI was used for coding purposes, although the authors inspected and are responsible for all details. AI was also used for generating irrelevant context in some settings, as described in the appendix.

REPRODUCIBILITY STATEMENT

We provide training and evaluation details within this work in the Appendix. Upon acceptance, all code and evaluation sets used within the paper will be open-sourced.

REFERENCES

Martín Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019. URL http://arxiv.org/abs/1907.02893.

Bradley C. A. Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *CoRR*, abs/2407.21787, 2024. doi: 10.48550/ARXIV.2407.21787. URL https://doi.org/10.48550/arXiv.2407.21787.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

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. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/ARXIV.2210.11416. URL https://doi.org/10.48550/arXiv.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. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. Gsm-plus: A comprehensive benchmark for evaluating the robustness of llms as mathematical problem solvers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 2961–2984. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.163. URL https://doi.org/10.18653/v1/2024.acl-long.163.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 20275–20321. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1025. URL https://doi.org/10.18653/v1/2025.emnlp-main.1025.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114, 2021. URL https://arxiv.org/abs/2112.00114.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 2020. URL http://proceedings.mlr.press/v119/sagawa20a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. URL https://doi.org/10.48550/arXiv.2402.03300.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa

Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a. URL https://openreview.net/forum?id=gEZrGCozdqR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Minglai Yang, Ethan Huang, Liang Zhang, Mihai Surdeanu, William Yang Wang, and Liangming Pan. How is LLM reasoning distracted by irrelevant context? an analysis using a controlled benchmark. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 13329–13347. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.674. URL https://doi.org/10.18653/v1/2025.emnlp-main.674.

Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Tn5B6Udq3E.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Second Conference on Language Modeling*, 2025.

APPENDIX

A EVALUATION SETS CREATION

We provide additional details to the evaluation sets described in the main text.

Extended GSM-IC. The original GSM-IC dataset (Shi et al., 2023) comes with many distractor sentence templates with slots for role and number. In our Extended GSM-IC test set, for each test sample from the original GSM8k dataset, we choose one sentence template, number and role and insert it before the final question sentence of that sample, similar to the GSM-IC dataset. Otherwise, the distractor is prepended to the whole question.

Irrelevance Auditing. To ensure that sentences inserted are irrelevant, we perform a comprehensive audit. We define a distractor sentence as irrelevant if its entities, objects, and quantities are not required to derive the answer to the original problem and it does not establish a relation that changes the original arithmetic computation. We use POS/NER-assisted filtering to differentiate between object/item overlaps from shared names, stopwords and generic temporal or unit terms. During our audit, we found that 52.01% of all samples had no overlap at all, 43.37% of all samples only shared a name, 4.17% of samples shared generic expressions such as day, week, or hour, and 0.45% of samples were flagged for concrete item/entity/object overlaps. The last category was manually reviewed to ensure that there was no overlap that was relevant to the arithmetic solution. We note that name overlap established only entity identity but not arithmetic relevance, since the IC's objects and quantities remain disconnected from those in the main question required for correct answering. As such, name only overlaps make it harder for the model to identify a sentence as truly irrelevant. All datasets will be open-sourced for the community to use and inspect.

Note that we do not use synthetic data generators as in (Yang et al., 2025) as such samples contain little diversity. Additionally, their generated datasets have been shown to contain instances of incorrect arithmetic and thus, we stick to the original GSM8k train and test sets with carefully audited distractors.

Train and Test Sourced Distractor Sets. For each test sample from the original GSM8k, we introduce the irrelevant context by sourcing them from the original GSM8k train and test set samples. From each of these, we split the source question into sentences and phrases based on relevant punctuations. We ensure that each distractor has at least 4 words. For the Test Sourced Distractor Set, we ensure that the source distractor is not drawn from the same question. We also ensure not to introduce any questions as distractors. The 2- and 3-sentence distractors are formed in a similar fashion but with multiple distractors scattered through the question.

We note that, in none of the samples, the distractor starts in the middle of a relevant sentence or phrase. We perform similar auditing to ensure irrelevance as in the case of the Extended GSM-IC setup.

Number and Operation Cue Distractor Set. Here, we sample a pool of numbers and operation cues from the standard GSM8k train and test sets and scatter them across questions, even in the middle of sentences. Per question, we introduce 5 distractors.

MATH500 with IC. Distractors/ICs were sourced from a set of eight templates encompassing the following categories—triangle, rectangle, polynomial, divisibility, counting, sequence, data table, and equivalent fractions. The sentences did not contain names, pronouns referring to the original problem and also did not contain shared variable symbols. The eight categories were put into 7 domains—geometry, algebra, number theory, counting, sequences, data/statistics, and fractions. For each example within the MATH500 evaluation set, we exclude any distractor domain whose subject or domain-specific keywords occur in the original problem, gold solution or answer. We additionally blocked distractor domains based on the original MATH subject labels Hendrycks et al. (2021), even if no domain keyword was detected.

B DISTRACTOR AUGMENTED FINE-TUNING SETS CREATION

We provide additional details to the distractor-augmented fine-tuning sets described in the main text.

For the Train and Test sources distractors, we follow the same approach as in the Train and Test Sourced Distractor Evaluation Sets. For Non-numeric, we ensure to exclude any numbers. For Out-of-domain, we ask GPT-5.4 to generate only 101 non-arithmetic sentences. We present representative examples in Sec. 4.1. When training with a mixture of distractors (Sec. 4.3), we scatter numbers and operation cue distractors randomly across the question.

C Ablations

In Sec. 3.1, we establish the learning of shortcut correlations by removing components from the distractors. Below, we discuss the heuristics used to identify the respective components:

- **Numbers.** These are identified through regular expression matching.
- **Operation Cues.** We use a fixed pool of operation cues extracted from the GSM8k train set: `left, remain, remaining, each, every, total, altogether, split, share, times, double, doubles, half, twice, more, less, fewer, difference, average, percent, per, rate, ratio, sum, product.`
  Operation cues are words often explicitly used to perform an arithmetic operation.
- **Names and Objects.** For identification of names and objects, we make use of spaCy, which is an open source python library for natural language processing.

C.1 Experimental Details

Statistical Significance. Error bars indicate ± 1 standard deviation across results from 3 random seeds.

CoT Length Computation. We compute the length of CoT by tokenizing it with each model's respective tokenizer and counting the total number of tokens.

D Sample Difficulty

In our experiments in Sec. 3.2, we show that samples containing long CoTs encourage a model to rely on irrelevant context. One might speculate that more difficult samples may have longer CoTs, and thus, sample difficulty can be a contributing factor to irrelevant context reliance. To confirm that sample difficulty does not play a notable role in irrelevant context reliance, we prune by sample difficulty instead of CoT length. We compute sample difficulty as follows: Train the model for 1 epoch and Compute average token loss. Samples with high average token loss are ranked as more difficult while samples with low average token loss are ranked as less difficult. We then prune samples by difficulty as follows:

- `Easy pruned:` We prune samples with the lowest loss (least difficult) as we move towards samples with the highest loss (most difficult).
- `Difficult pruned:` We prune samples with the highest loss (most difficult) as we move towards samples with the lowest loss (least difficult).

We observe that pruning by difficulty does not produce the same patterns as pruning by CoT length. The results do not convey any significant trend, as shown in Fig. 9. Hence, we conclude that sample difficulty does not impact irrelevant context reliance.

E Comparing Fine-tuning with IC to Prompting

In this section, we show how models behave when fine-tuned with ICs as shown in Fig. 7 versus when they have simply been prompted to ignore the irrelevant context. For the latter, we add "Please Ignore Irrelevant Information" to the original prompt. We observe that simply prompting does not offer the same benefits compared to fine-tuning with IC (Fig. 10). In our Gemma3-4B setup, it offers almost no improvements. This implies that adding IC within the fine-tuning set provides training signals to overcome IC.

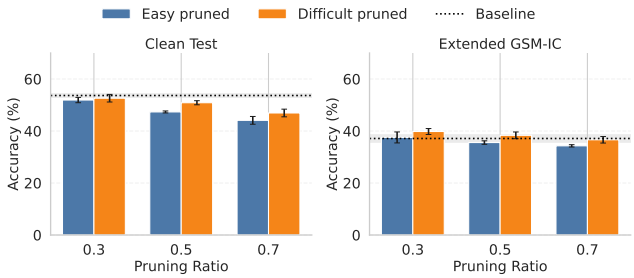


Figure 9: Pruning by sample difficulty does *not* produce the same patterns as pruning by CoT length. Results shown with Gemma3-4B.

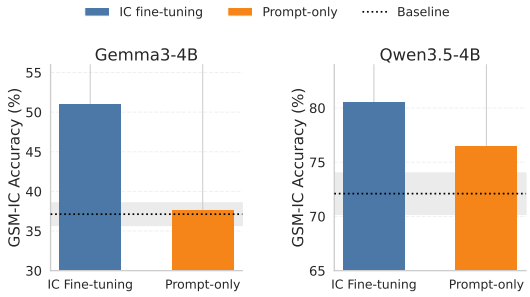


Figure 10: Prompting the model to overcome IC does not offer the same benefits as fine-tuning the model on data with IC. Performance shown on the Extended GSM-IC dataset.

F Training Details

We utilize recent open-weight models in our experiments. All Qwen3.5-4B (Qwen Team, 2026) models were fine-tuned for 3 epochs using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1e-4 with cosine decay and weight decay 0.1. We utilize LoRA (Hu et al., 2022) fine-tuning with rank 64, alpha 16, and dropout 0.1. All Gemma3-4B (Team et al., 2025) models were fine-tuned with LoRA for 5 epochs using AdamW with a learning rate of 5e-4 with cosine decay and weight decay 0.1. We utilize LoRA fine-tuning with rank 8, alpha 16, and dropout 0.1. Pre-trained large language models are fine-tuned on reasoning tasks through two broad techniques—Supervised fine-tuning (Wei et al., 2022a; Chung et al., 2022) and reinforcement learning (Shao et al., 2024; Zeng et al., 2025). In this work, we focus on supervised fine-tuning, as it is the de facto standard for fine-tuning, given it is notably cheaper and does not require verification of outputs for reward/credit assignment.

G Compute Report

All experiments in this work were performed on a machine with 2 NVIDIA RTX 4090s. Additionally, time taken to fine-tune any model on any given dataset is at most 2 hours, making our experiments easily reproducible for the community.

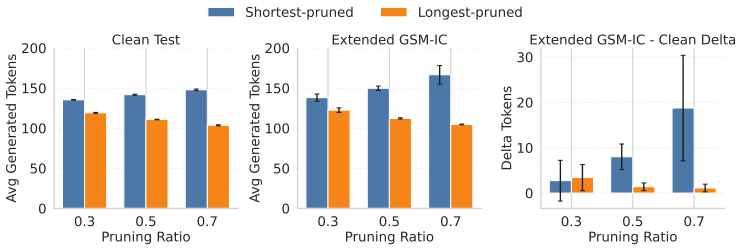


Figure 11: Test time impact of two different data pruning strategies in train time. Long CoT Supervision (`Shortest Pruned`) teaches the model to write verbose reasoning steps during test time, encouraging it to draw in irrelevant information. Results shown with Qwen3.5-4B.