The Feynman Trap: When Chain-of-Thought Reasoning Undermines Correct Intuitions in Language Models

We systematically quantify a phenomenon where LLMs flip correct zero-shot answers to incorrect ones when prompted to reason step-by-step, and characterize when this cost outweighs the benefit across models and task domains.

AUTHORS
Anonymous

AFFILIATIONS
Anonymous

SUBMITTED
Not submitted yet.

Contents

Introduction

Related Work

Experimental Setup

The Feynman Trap: CoT-Induced Regression

Cross-Domain Analysis

Understanding the Mechanism

Mitigation via Self-Consistency

Discussion and Limitations

Conclusion

Introduction

Chain-of-thought (CoT) prompting [1] has become the default strategy for improving reasoning in large language models (LLMs). By instructing models to "think step by step," CoT elicits intermediate reasoning that substantially boosts performance on mathematical, logical, and commonsense tasks. This success has cemented a widespread assumption: *more reasoning is always better*.

We challenge this assumption by identifying and systematically quantifying a phenomenon we term the **Feynman Trap** — named after Richard Feynman's observation that *"the first principle is that you must not fool yourself, and you are the easiest person to fool."* Specifically, we find that LLMs which answer correctly under zero-shot (ZS) conditions frequently produce *wrong* answers when prompted to reason step-by-step. The reasoning process appears to corrupt correct intuitions.

This is not merely a matter of CoT failing to help on some questions. We document a specific, directional failure mode: answers that were correct without reasoning become incorrect *under* reasoning. Across four 7B-class models on GSM8K [2], 9.5–88% of zero-shot-correct answers are flipped to incorrect by CoT (after correcting for extraction artifacts), with the flip rate inversely associated with model capability. On CoQA conversational QA [3] (n=500, evaluated via F1 overlap), CoT produces a net accuracy loss for all four models (−3.6 to −26.8 percentage points), though the constrained yes/no subset shows a more nuanced, capability-dependent pattern.

Our contributions are:

1. **Systematic quantification** of CoT-induced correct-to-incorrect (C→I) flips across 4 models and 2 task domains, with corrected flip rates that account for extraction artifacts.
2. **Temporal analysis** showing models commit to incorrect answers in the first half of the reasoning chain (median 64–145 tokens).
3. **Counterfactual controls** showing that a format-free "think briefly" prompt produces comparable flip rates to structured CoT, suggesting the flips are not solely due to output formatting or extraction artifacts.
4. **Cross-domain divergence**: CoT is net positive on math but net negative on conversational text QA (full CoQA, F1 evaluation), with a capability-dependent pattern on constrained binary questions.
5. **Mitigation analysis**: Self-consistency (SC@5) rescues 9–67% of flips, with rescue rate positively associated with model capability.

Related Work

Chain-of-thought prompting and variants. Wei et al. [1] demonstrated that CoT prompting substantially improves reasoning in LLMs, an approach extended by zero-shot CoT [4], self-consistency decoding [5], and scratchpad-based reasoning [6]. Subsequent work introduced structured decomposition strategies: least-to-most prompting [7], tree of thoughts [8], plan-and-solve prompting [9], and complexity-based prompting [10]. Program-of-thoughts [11] disentangles computation from reasoning by generating executable code. These works uniformly report aggregate accuracy *gains* from CoT without examining per-example costs — our work complements them by quantifying the directional C→I flip that these aggregate metrics obscure.

Reasoning quality and verification. Lightman et al. [12] introduced process reward models that verify each reasoning step, finding that step-level supervision outperforms outcome-level supervision. Zelikman et al. [13] bootstrap reasoning via STaR, iteratively improving rationales. Saparov and He [14] showed that language models are "greedy reasoners" that commit to locally plausible steps without global planning — a finding consistent with our temporal analysis showing early commitment to errors. Wang and Zhou [15] demonstrated that CoT reasoning can emerge without explicit prompting via decoding strategies, suggesting the phenomenon is not purely prompt-induced.

Faithfulness and reliability of reasoning. Turpin et al. [16] showed that CoT explanations can be unfaithful — the stated reasoning may not reflect the model's actual computation. Lanham et al. [17] measured faithfulness by truncating and corrupting chains, finding that early tokens often suffice for the final answer. Shi et al. [18] demonstrated that LLMs are easily distracted by irrelevant context in reasoning tasks, with CoT amplifying rather than mitigating this vulnerability. Huang et al. [19] showed that LLMs cannot reliably self-correct reasoning without external feedback — consistent with our finding that the Verify condition (self-checking with the ZS answer) fails to rescue flips. Together, these works suggest that CoT chains are neither faithful representations of model computation nor reliable targets for self-correction.

Overthinking and reasoning costs. Several concurrent works examine reasoning costs. Chen et al. [20] revisit overthinking in long CoT, focusing on compute efficiency. Li et al. [21] show reasoning can hurt inductive abilities. Wu et al. [22] study overthinking in reasoning-specialized LLMs. Madaan et al. [23] proposed iterative self-refinement, though subsequent work has questioned its effectiveness without external signals. Our work differs in three ways: (1) we focus specifically on the *directional* C→I flip rather than aggregate degradation, (2) we provide counterfactual prompt controls suggesting flips are not solely due to formatting artifacts, and (3) we characterize the cross-domain asymmetry (CoT helps math, hurts conversational text QA).

Emergent capabilities and scaling. Wei et al. [24] documented emergent abilities in large language models, including the observation that CoT benefits scale with model size — an important context for our finding that the Feynman Trap is more severe in weaker models. Brown et al. [25] established the few-shot learning paradigm that CoT builds upon. Zhang et al. [26] extended CoT to multimodal settings, showing similar reasoning benefits and potential failure modes across modalities.

Experimental Setup

Models. We evaluate four 7B-class models spanning different architectures and training approaches: Qwen2.5-7B [27], Llama-3.1-8B [28], Llama-2-7B-chat [29], and Mistral-7B-v0.2 [30]. These models represent a natural capability spectrum at the 7B scale: Qwen2.5-7B is the strongest (75.4% CoT accuracy on GSM8K) and Mistral-7B the weakest (7.2%).

Datasets. We use two benchmarks:

- **GSM8K** [2]: 1,319 grade-school math word problems with numerical answers. Structured task with a single correct answer per question.

- **CoQA** [3]: 500 conversational question-answer pairs requiring reading comprehension. Open-ended text answers with evaluation via F1 overlap.

Protocol. For each model × dataset combination, we collect:

1. **Zero-shot (ZS)** responses: direct answer without reasoning.
2. **CoT responses**: "Let's think step by step" prompt with model-appropriate answer extraction (e.g., `#### [number]` for GSM8K).
3. **Self-consistency (SC@5)**: 5 independent CoT samples at temperature 0.7, majority vote.
4. **Counterfactual conditions** (GSM8K only): "Think briefly" and "Verify" prompts.

A **C→I flip** is defined as a question where the model answers correctly under ZS but incorrectly under CoT. All experiments use greedy decoding (temperature 0) for ZS and CoT, with generation on 4×A100-40GB GPUs.

Statistical rigor. We report McNemar's paired test [31] for ZS vs CoT comparisons, Cohen's *h* effect sizes [32], bootstrap 95% confidence intervals (10K resamples), and permutation tests (10K iterations) for base-rate validation.

The Feynman Trap: CoT-Induced Regression

Main Results

The table below summarizes the core finding across all four models on GSM8K (n=1,319 per model). We report both raw flip rates and corrected rates. Correction is performed via automated suffix-stripping and re-extraction applied to *all* C→I flips (not sampled): for each flip, we strip known model-appended suffixes (e.g., "You are an AI assistant" for Qwen) and re-extract the numerical answer. If re-extraction recovers the correct answer, the flip is reclassified as an extraction artifact. This procedure is deterministic and applied exhaustively to all flips.

Model	ZS (%)	CoT (%)	Δ	Raw C→I	Raw Flip Rate	Corrected C→I	Corrected Flip Rate	McNemar χ^2	Cohen's *h*
Qwen2.5-7B	18.3	75.4	+57.1	39	16.2%	23	**9.5%** [95% CI: 5.8, 13.9]	681.5***	+1.22
Llama-3.1-8B	14.6	23.4	+8.9	131	68.2%	131	**68.2%** [61.5, 74.5]	35.5***	+0.23
Llama-2-7B	4.2	15.5	+11.3	46	83.6%	41	**74.5%** [63.6, 85.5]	90.9***	+0.40
Mistral-7B	4.5	7.2	+2.7	52	88.1%	52	**88.1%** [79.7, 94.9]	8.8**	+0.12

*** $p < 0.001$, ** $p < 0.01$. Raw Flip Rate = C→I / (ZS correct). Corrected Flip Rate excludes extraction artifacts identified by automated re-extraction on all flips; 95% bootstrap CIs in brackets. McNemar tests use raw counts.

CoT is net positive for all four models — the I→C (incorrect→correct) benefit dominates. Yet the C→I cost is substantial: after automated correction for known extraction artifacts (a conservative lower bound on true artifact rates), **Mistral still flips 88% of its zero-shot-correct answers when asked to reason**, and even Qwen, the strongest model, loses 9.5% of its correct answers to reasoning errors under CoT.

Figure 1. Answer flow from zero-shot to chain-of-thought. Select a model to see how each answer transitions. The red stream shows the Feynman Trap: correct answers that become incorrect under CoT. Hover over flows for counts.

We observe an inverse association between capability and flip rate across these four models ($r = -0.98$, though we note this correlation is computed over only 4 points and should be treated as an exploratory observation rather than a robust scaling law). The pattern is suggestive: stronger models may have more robust internal representations that survive the perturbation introduced by step-by-step reasoning, while weaker models' fragile correct answers are more easily overwritten. Confirming this trend would require evaluation across a wider range of model scales.

Base-Rate Validation

A natural concern is that zero-shot correct answers at 4–18% accuracy could be lucky guesses. Permutation tests (10K iterations) confirm this is not the case:

Model	ZS Accuracy	Random Baseline	Ratio	*p*-value
Qwen2.5-7B	18.3%	1.60%	11.4×	< 0.01
Llama-3.1-8B	14.6%	1.24%	11.8×	< 0.01
Mistral-7B	4.5%	1.58%	2.8×	< 0.01
Llama-2-7B	4.2%	0.80%	5.2×	< 0.01

All models exceed the 99th percentile of the random distribution, confirming genuine zero-shot mathematical competence. The ZS-correct answers that CoT subsequently flips represent real knowledge, not noise.

Extraction Confound

The corrected flip rates above are computed via automated re-

extraction applied to all flips. To independently validate this automated procedure and characterize the remaining genuine errors, we conduct a human audit of 100 randomly sampled flips (25 per model). The audit reveals two distinct failure modes:

Category	Count	Proportion
Extraction artifact	56	56%
Genuine reasoning error	30	30%
Grabbed number from question	13	13%
Near miss (within 5%)	1	1%

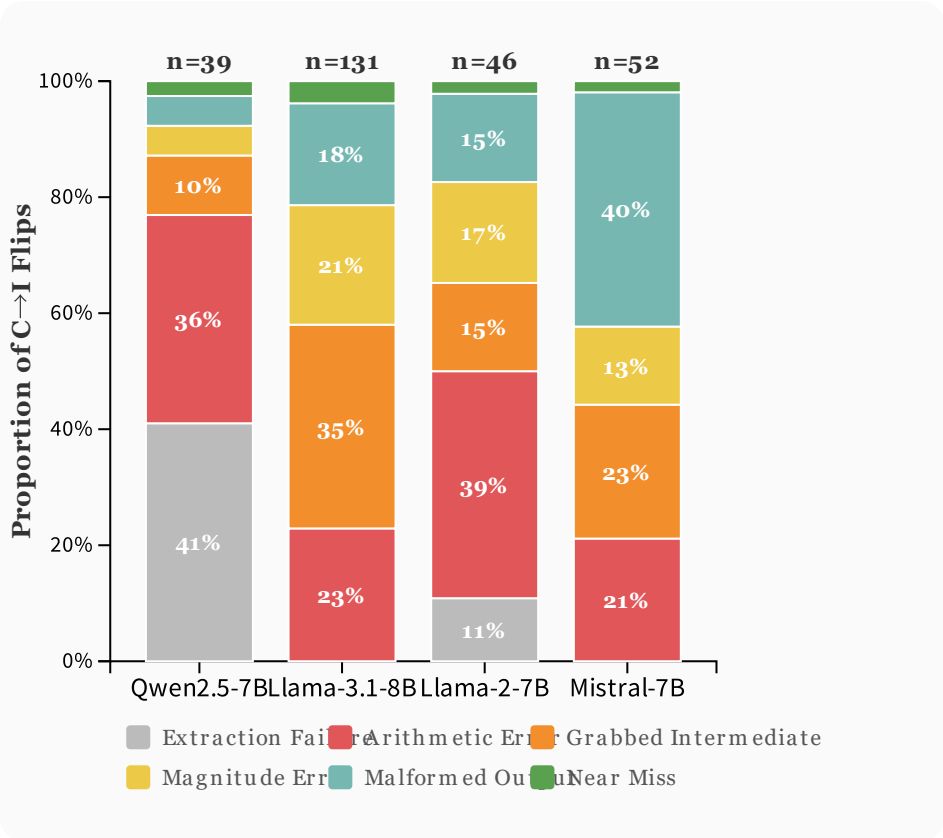

Figure 2. Error taxonomy of C→I flips by model. Hover over segments for details. Extraction failures (gray) represent pipeline artifacts where the model's reasoning was correct; genuine errors (colored) represent true reasoning corruption. The proportion of extraction failures decreases with weaker models.

Extraction failures occur when the model reaches the correct answer but the extraction pipeline fails — typically because the model appends system-prompt text or wraps answers in non-standard formatting. These are pipeline artifacts, not reasoning failures. The automated correction procedure (applied to all flips) reduces Qwen's flip count from 39 to 23 (corrected rate: 9.5%) and Llama-2's from 46 to 41 (corrected rate: 74.5%). For Llama-3.1 and Mistral, no extraction artifacts were detected by automated re-extraction. The human audit finds a higher artifact rate (56%) than the automated procedure (21/268 = 7.8%) for two reasons: (1) the audit samples 25 flips per model uniformly, overweighting Qwen (which has 16/39 = 41% automated artifacts but only 39 total flips vs. Llama-3.1's 131), and (2) the human audit uses a broader artifact definition — annotators classify any flip where the CoT reasoning arrived at the correct answer as an extraction artifact, catching cases that the automated suffix-stripping rule misses (e.g., formatting variations beyond known suffixes). The automated procedure is conservative by design; the audit provides an upper bound. The true rate of genuine reasoning corruption likely lies between the automated estimate and the audit estimate — the corrected flip rates we report should be interpreted as upper bounds on genuine reasoning flips, not exact counts.

Genuine reasoning errors identified in the human audit include arithmetic mistakes (39% of Llama-2's audited flips), grabbing intermediate values from the problem text (35% of Llama-3.1), and malformed outputs (40% of Mistral). We report corrected flip rates with bootstrap 95% CIs as primary throughout.

To make these errors concrete, the following interactive explorer shows real flip examples from our experiments. Each card shows a GSM8K question where the model answered correctly without reasoning but produced the wrong answer after step-by-step thinking. The error in the CoT chain is highlighted in red.

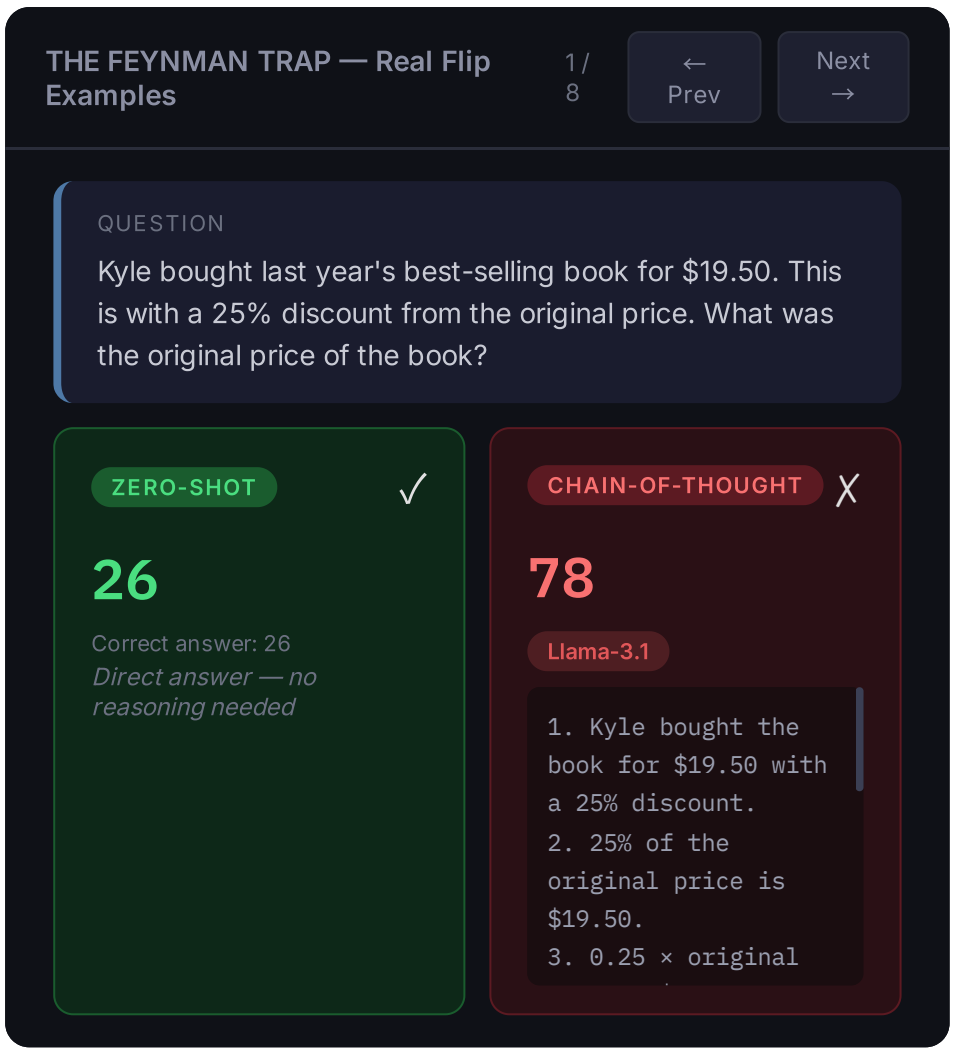

Figure 3. Real flip examples from our experiments. Use arrow keys or buttons to browse. Each card shows a question where the model knew the answer without reasoning (green, left) but produced a wrong answer after step-by-step thinking (red, right). The reasoning chain reveals the specific error.

Cross-Domain Analysis

A critical question is whether the Feynman Trap is specific to math or generalizes across domains. We evaluate all four models on CoQA conversational question answering.

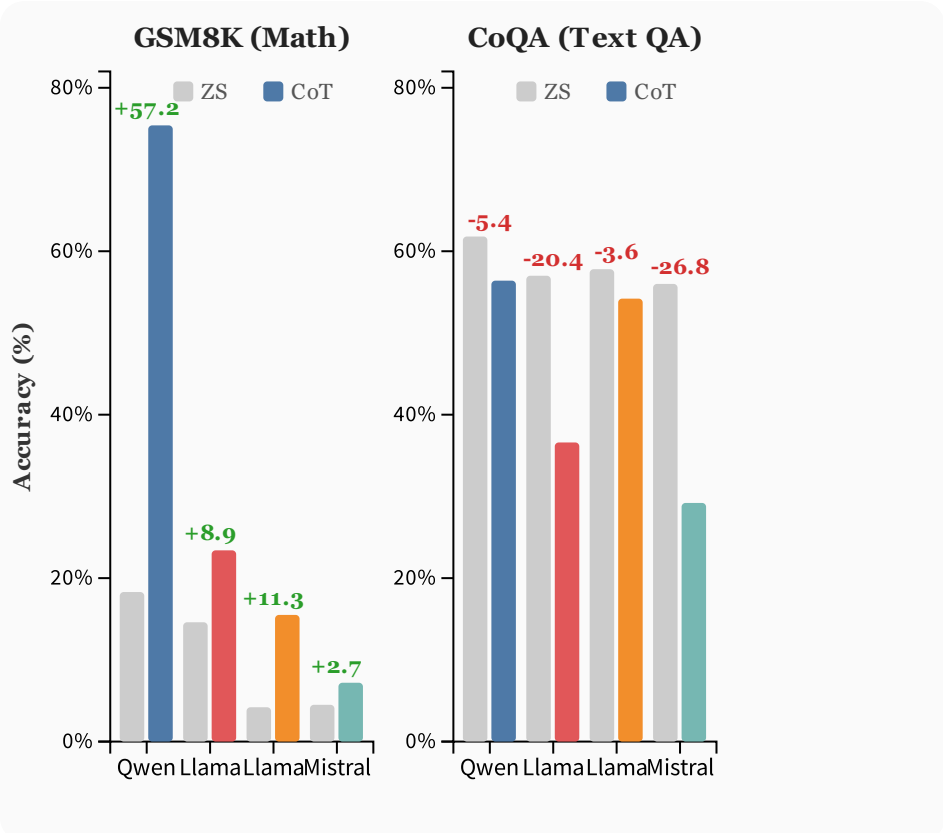

Figure 4. ZS vs. CoT accuracy on GSM8K (math) and CoQA (conversational QA, F1 evaluation). Hover for details. Green annotations indicate CoT improvement; red indicates degradation. CoT is net positive on math but net negative on CoQA (F1 evaluation) for all four models.

Model	CoQA ZS (%)	CoQA CoT (%)	Δ	C→I Flips	Flip Rate
Qwen2.5-7B	61.8	56.4	**−5.4**	51	16.5%
Llama-2-7B	57.8	54.2	**−3.6**	55	19.0%
Llama-3.1-8B	57.0	36.6	**−20.4**	131	46.0%
Mistral-7B	56.0	29.2	**−26.8**	164	58.6%

On the full CoQA validation set (n=500, evaluated via F1 overlap), **CoT produces a net accuracy loss for all four models.** Models already answer correctly 56–62% of the time via direct pattern matching; step-by-step reasoning degrades performance. The degradation ranges from mild (Qwen, −5.4%) to severe (Mistral, −26.8%). Note that this set includes both open-ended and constrained (yes/no) questions — the constrained binary subset below reveals a more nuanced, capability-dependent pattern.

Yes/No Subset: Zero Extraction Ambiguity

To eliminate any extraction confound, we evaluate on the 110 yes/no questions from CoQA where answers are constrained to binary choices:

Model	ZS (%)	CoT (%)	Δ	C→I Flips	Flip Rate
Qwen2.5-7B	86.4	88.2	+1.8	4	4.2%
Llama-2-7B	79.1	86.4	+7.3	7	8.0%
Llama-3.1-8B	74.5	36.4	**−38.1**	47	57.3%
Mistral-7B	70.0	30.9	**−39.1**	56	72.7%

The Yes/No subset reveals a descriptive split: **stronger models (Qwen, Llama-2) show small positive trends under CoT** (+1.8%, +7.3%), though neither reaches statistical significance by exact sign test (p=0.75, p=0.13 respectively given the small discordant counts). In contrast, **weaker models (Llama-3.1, Mistral) lose 38–39 percentage points**. This nuances the cross-domain finding: the Feynman Trap on text QA appears capability-dependent rather than universal. The fact that weaker models lose dramatically on binary questions with zero extraction ambiguity confirms the Feynman Trap is a genuine reasoning phenomenon for these models, not an evaluation artifact.

Understanding the Mechanism

Temporal Analysis: When Do Flips Happen?

We use progressive CoT checkpointing — extracting answers at token positions [16, 32, 48, 64, 96, 128, 192, 256, 384, 512] — with a **committed-flip criterion** defined as follows: for each C→I flip, we identify the earliest checkpoint t at which the extracted answer matches the model's final (incorrect) CoT answer, and this match persists at all subsequent checkpoints. This requires not just a one-time match but sustained commitment. Checkpoints below 32 tokens are excluded from commitment analysis due to high extraction noise (validated by checkpoint-to-final agreement rates of only 2–10% at ≤16 tokens).

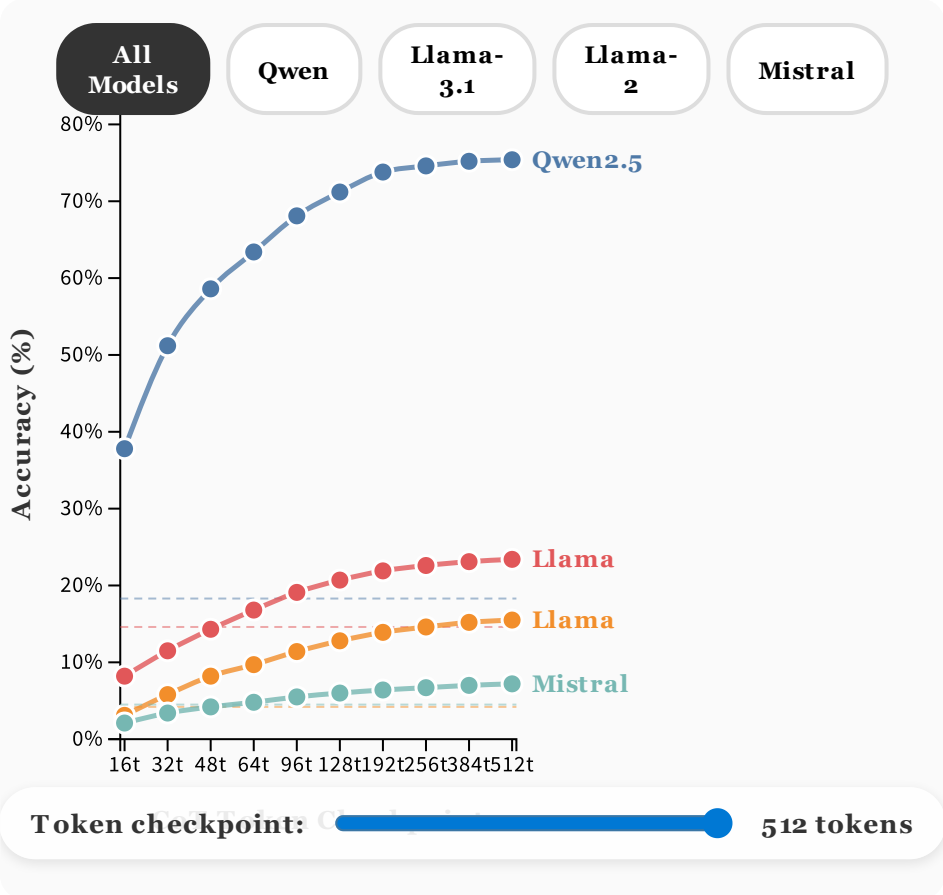

Figure 5. Progressive CoT accuracy at each token checkpoint. Use buttons to isolate individual models; use the slider to truncate the visible chain length. Dashed lines show ZS baselines. Models commit to incorrect answers in the first half of their reasoning chains.

Model	Median Committed Flip	≤64 tokens	≤128 tokens	Mean Chain Length
Llama-3.1-8B	64 tokens	52%	69%	554 tokens
Qwen2.5-7B	128 tokens	13%	51%	197 tokens
Mistral-7B	128 tokens	42%	69%	700 tokens
Llama-2-7B	145 tokens	15%	48%	197 tokens

Models commit to incorrect answers at median 64–145 tokens, well before reasoning completes. For Llama-3.1, 52% of committed flips occur within the first 64 tokens of a 554-token chain. For these flipped examples, the remaining reasoning serves to elaborate on an already-committed incorrect path rather than recovering the correct answer.

Counterfactual Prompt Controls

A key question is whether flips are primarily associated with the *reasoning process itself* or with CoT-specific formatting and extraction artifacts. While a fully controlled causal experiment would require holding output format and length constant while varying only reasoning depth — which is difficult with autoregressive LLMs — we approximate this by testing four prompt conditions on GSM8K that vary in reasoning structure:

- **ZS**: Direct answer (baseline)
- **CoT**: Full step-by-step with structured format
- **Brief**: "Think briefly about this problem, then give your answer" (reasoning without rigid structure)
- **Verify**: "Someone suggested the answer is [ZS answer]. Verify step by step." (anchor + reasoning)

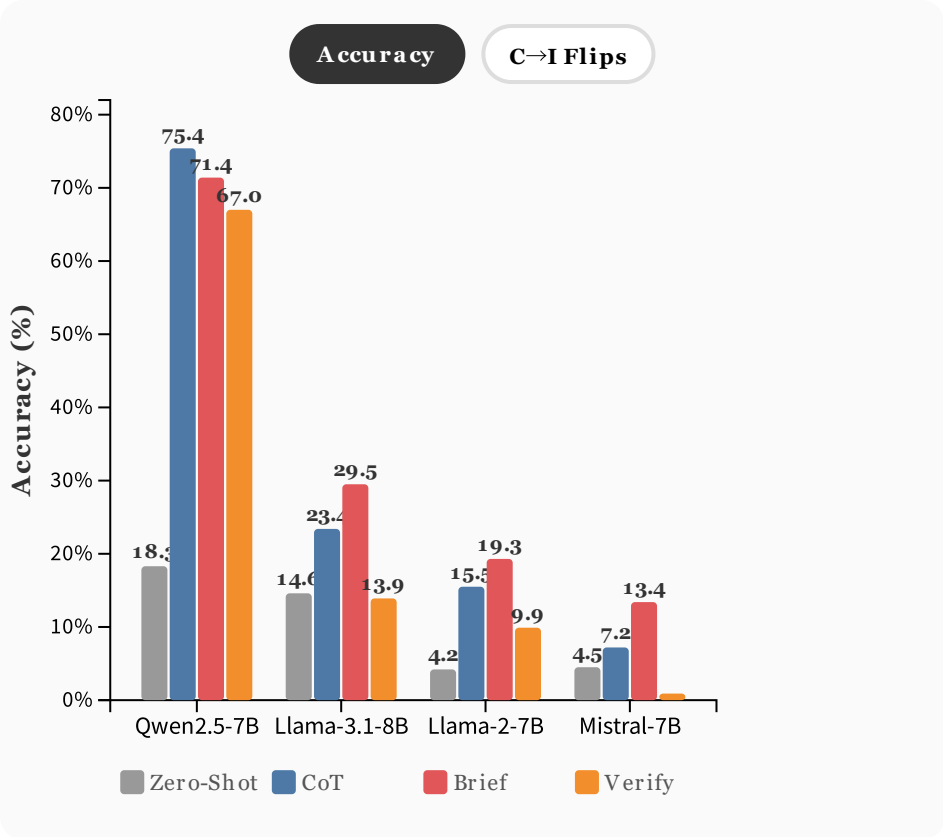

Figure 6. Accuracy and flip counts under four prompt conditions. Toggle between "Accuracy" and "C→I Flips" views. Brief reasoning produces comparable flip rates to full CoT, suggesting flips are not solely due to CoT formatting artifacts.

Model	ZS (%)	CoT (%)	Brief (%)	Verify (%)	CoT Flips	Brief Flips
Qwen2.5-7B	18.3	75.4	71.4	67.0	39	35
Llama-3.1-8B	14.6	23.4	29.5	13.9	131	106
Llama-2-7B	4.2	15.5	19.3	9.9	46	38
Mistral-7B	4.5	7.2	13.4	0.9	52	33

Three findings emerge:

1. **Brief reasoning causes flips too** (33–106 vs. CoT's 39–131). Since the Brief condition uses no structured format or extraction markers (e.g., `####`), this suggests the flips are not solely attributable to the structured CoT format or extraction pipeline. However, we cannot fully rule out that verbosity, prompt wording, or output length contribute independently.

2. **Brief outperforms CoT for weaker models**: Llama-3.1 gains +6.1%, Llama-2 gains +3.8%, Mistral gains +6.2% with brief prompting over full CoT. The rigid step-by-step format adds overhead that hurts weaker models more than it helps. Only Qwen (the strongest) benefits from full CoT structure.

3. **Verify (anchoring) degrades overall accuracy.** Providing the ZS answer and asking to verify introduces *new* errors (Mistral drops to 0.9%). Models cannot reliably use their own zero-shot answers as verification anchors.

Mitigation via Self-Consistency

Self-consistency [5] samples multiple CoT paths and takes a majority vote, potentially recovering correct answers when the flip is stochastic rather than systematic.

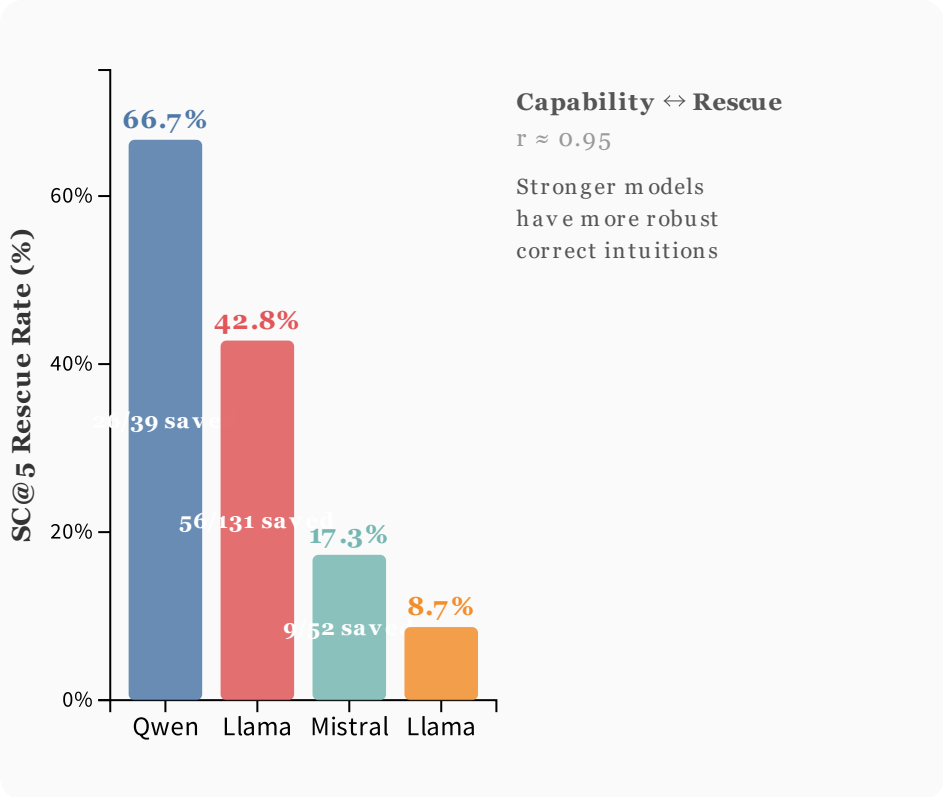

Figure 7. SC@5 rescue rate by model. Hover for detailed flip counts and accuracy changes. The rescue rate shows a positive association with model capability ($r \approx 0.95$, 4 models; exploratory).

Model	CoT (%)	SC@5 (%)	Δ	CoT Flips	SC Flips	Rescues	Rescue Rate [95% CI]
Qwen2.5-7B	75.4	77.6	+2.1	39	25	26	**66.7%** [51.0, 80.0]
Llama-3.1-8B	23.4	32.4	+8.9	131	101	56	**42.8%** [34.4, 51.1]
Mistral-7B	7.2	10.6	+3.4	52	45	9	17.3% [8.2, 29.4]
Llama-2-7B	15.5	15.8	+0.4	46	45	4	8.7% [2.4, 19.6]

Rescue Rate 95% CIs computed via binomial proportion confidence intervals (Clopper-Pearson exact method).

SC@5 rescue rate shows a positive association with model capability ($r \approx 0.95$ over 4 models; exploratory). **Qwen recovers 67% of its flips** [95% CI: 51–80%] — its correct intuition is robust enough to survive in the majority of 5 sampled reasoning paths. Llama-2, the model with the lowest SC rescue rate, recovers only 9% [95% CI: 2–20%] — the same incorrect reasoning recurs systematically across samples. (Note: Mistral is the weakest model by CoT accuracy but has a higher rescue rate than Llama-2, suggesting rescue rate is not strictly monotonic with capability.)

SC struggles on text QA. We ran SC@5 on CoQA for two models (Qwen and Llama-2): SC@5 accuracy dropped to 9.2% and 10.6% respectively (vs. CoT 56.4% and 54.2%), because majority vote fragments across diverse text phrasings of the same answer. With 5 samples producing 5 different surface-form answers (e.g., "New York", "in New York", "New York City"), no single phrasing achieves a majority. SC's rescue mechanism requires a single canonical correct answer (as in math) to aggregate votes effectively.

Discussion and Limitations

Why Does Reasoning Corrupt Correct Answers?

Our results suggest that CoT prompting introduces a *reasoning perturbation* that can override correct direct-retrieval answers. The model's zero-shot answer reflects a holistic pattern-match over the input; CoT forces a serial decomposition that creates opportunities for error propagation. Once an early step goes wrong, the remaining chain follows the erroneous path — consistent with our temporal analysis showing early commitment to errors.

The counterfactual experiments are consistent with this interpretation: the Brief condition ("think briefly") — which lacks structured step-by-step formatting — produces comparable flip rates to full CoT, suggesting the flips are not solely due to format-specific artifacts, though the reasoning process likely contributes. We acknowledge that these prompt controls do not constitute a fully controlled causal experiment, as they also vary verbosity, output length, and prompt wording. A stronger causal identification would require interventions that hold these factors constant while varying only reasoning depth.

When Should Practitioners Use CoT?

The interactive dashboard below lets readers explore the cost-benefit tradeoff for any model × task combination in our study, comparing all prompt strategies side-by-side.

Figure 8. Interactive dashboard: select a model and task to see the full cost-benefit picture. Compares accuracy, flip rates, and all prompt strategies. The "BEST" tag highlights the optimal prompting strategy for each configuration.

Our cross-domain results suggest a nuanced prescription:

- **Structured tasks** (math, logic): CoT provides large net gains despite the flip cost. SC can partially mitigate flips for capable models.
- **Conversational text QA** (CoQA): CoT is net negative under F1 evaluation for all four models tested, especially for weaker models. However, stronger models show non-significant

positive trends on constrained binary questions. Direct prompting may be preferable for weaker models on such tasks.

- **Weaker models on math**: Brief prompting outperforms full CoT on GSM8K for 3 of 4 models, suggesting that rigid step-by-step structure adds overhead that can exceed the reasoning benefit at lower capability levels. (Brief was only tested on GSM8K.)

Limitations

Model scale and scaling claims. We study four 7B-class models. The observed inverse association between capability and flip rate ($r = -0.98$) is computed over only 4 data points and should be treated as an exploratory observation, not a robust scaling law. Confirming this trend requires evaluation across more model sizes. Similarly, the SC rescue rate correlation ($r \approx 0.95$) is exploratory.

Two task families. GSM8K and CoQA represent structured math and conversational text QA. Other task types (code generation, multi-step planning, scientific reasoning) may show different patterns.

Cross-domain nuance. The claim that "CoT hurts text QA" holds for all four models on the full CoQA set (F1 evaluation), but the Yes/No subset shows that stronger models (Qwen, Llama-2) show small positive trends under CoT (not statistically significant). The cross-domain finding is therefore evaluation-protocol-dependent and capability-dependent.

Extraction artifacts. Despite corrected flip rates and human audit (n=100), some residual extraction noise may remain. We foreground corrected rates and use the Yes/No CoQA subset (zero ambiguity) as an independent validation.

Prompt sensitivity. Our CoT prompt ("Let's think step by step") is one of many possible formulations. Different prompt wordings or few-shot examples could yield different flip rates. Our counterfactual controls vary the reasoning instruction but do not constitute a fully controlled experiment — they also vary verbosity and output format.

Counterfactual controls. The "Brief" and "Verify" conditions are suggestive but not clean causal manipulations. They show that format-free reasoning also causes flips, which is consistent with reasoning being the driver, but cannot fully isolate reasoning from correlated factors (prompt length, verbosity, attention distribution).

Conclusion

We identify and quantify the **Feynman Trap**: a systematic phenomenon where chain-of-thought prompting corrupts correct zero-shot answers in LLMs. Across four 7B-class models and two task domains, we find:

- **9.5–88%** of zero-shot-correct answers flip to incorrect under CoT (corrected for extraction artifacts), with an inverse association with model capability.
- CoT is net positive on math but **net negative on CoQA text QA** (F1 evaluation), though the constrained yes/no subset shows a capability-dependent pattern (not statistically significant for stronger models).
- Format-free brief reasoning produces comparable flip rates to structured CoT, suggesting the reasoning process contributes to

flips beyond format-specific artifacts.

- Models commit to errors **early** (median 64–145 tokens under our committed-flip criterion), then elaborate without self-correction.
- Self-consistency rescues **9–67%** of flips, with a generally positive but non-monotonic association with model capability.

These findings challenge the assumption that reasoning is uniformly beneficial and suggest that the choice to use CoT should be task- and model-dependent. We hope this work encourages the community to evaluate CoT not only by its aggregate gains but also by its per-example costs.

References

1. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**
   Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V. and Zhou, D., 2022. Advances in Neural Information Processing Systems, Vol 35.
2. **Training Verifiers to Solve Math Word Problems**
   Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. and Schulman, J., 2021. arXiv preprint arXiv:2110.14168.
3. **{C}o{QA}: A Conversational Question Answering Challenge** [link]
   Reddy, S., Chen, D. and Manning, C.D., 2019. Transactions of the Association for Computational Linguistics, Vol 7, pp. 249--266. MIT Press. DOI: 10.1162/tacl_a_00266
4. **Large Language Models are Zero-Shot Reasoners** [link]
   Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y., 2022. ICML 2022 Workshop on Knowledge Retrieval and Language Models.
5. **Self-Consistency Improves Chain of Thought Reasoning in Language Models**
   Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A. and Zhou, D., 2023. International Conference on Learning Representations.
6. **Show Your Work: Scratchpads for Intermediate Computation with Language Models**
   Nye, M., Andreassen, A.J., Gur-Ari, G., Michalewski, H., Austin, J., Biber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D. and others,, 2022. International Conference on Learning Representations.
7. **Least-to-Most Prompting Enables Complex Reasoning in Large Language Models**
   Zhou, D., Scharli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. and Chi, E., 2023. International Conference on Learning Representations.
8. **Tree of Thoughts: Deliberate Problem Solving with Large Language Models**
   Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T.L., Cao, Y. and Narasimhan, K., 2023. Advances in Neural Information Processing Systems, Vol 36.
9. **Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models**
   Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R.K. and Lim, E., 2023. Annual Meeting of the Association for Computational Linguistics.
10. **Complexity-Based Prompting for Multi-step Reasoning**
    Fu, Y., Peng, H., Sabharwal, A., Clark, P. and Khot, T., 2023. International Conference on Learning Representations.
11. **Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks**
    Chen, W., Ma, X., Wang, X. and Cohen, W.W., 2023. Transactions on Machine Learning Research.
12. **Let's Verify Step by Step**
    Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J.,

Schulman, J., Sutskever, I. and Cobbe, K., 2024. International Conference on Learning Representations.

13. **STaR: Bootstrapping Reasoning With Reasoning** [PDF]
Zelikman, E., Wu, Y., Mu, J. and Goodman, N., 2022. Advances in Neural Information Processing Systems, Vol 35, pp. 15476--15488. Curran Associates, Inc.

14. **Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought**
Saparov, A. and He, H., 2023. International Conference on Learning Representations.

15. **Chain-of-Thought Reasoning Without Prompting**
Wang, X. and Zhou, D., 2024. arXiv preprint arXiv:2402.10200.

16. **Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting**
Turpin, M., Michael, J., Perez, E. and Bowman, S.R., 2024. Advances in Neural Information Processing Systems.

17. **Measuring Faithfulness in Chain-of-Thought Reasoning**
Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J. and others,, 2023. arXiv preprint arXiv:2307.13702.

18. **Large Language Models Can Be Easily Distracted by Irrelevant Context**
Shi, F., Chen, X., Mishra, K., Scales, N., Dohan, D., Chi, E.H., Scharli, N. and Zhou, D., 2023. International Conference on Machine Learning.

19. **Large Language Models Cannot Self-Correct Reasoning Yet**
Huang, J., Chen, X., Mishra, S., Zheng, H.S., Yu, A.W., Song, X. and Zhou, D., 2024. International Conference on Learning Representations.

20. **Revisiting Overthinking in Long Chain-of-Thought from the Perspective of Self-Doubt** [PDF]
Peng, K., Ding, L., Ouyang, Y., Fang, M. and Tao, D., 2025.

21. **Reasoning Can Hurt the Inductive Abilities of Large Language Models** [PDF]
Jin, H., Zhang, P., Luo, M. and Wang, H., 2025.

22. **When More Thinking Hurts: Overthinking in LLM Test-Time Compute Scaling** [PDF]
Zhou, S., Ling, R., Chen, J., Wang, X., Fan, T. and Wang, H., 2026.

23. **Self-Refine: Iterative Refinement with Self-Feedback** [PDF]
Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A. and Clark, P., 2023. Advances in Neural Information Processing Systems, Vol 36, pp. 46534--46594. Curran Associates, Inc.

24. **Emergent Abilities of Large Language Models**
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J. and Fedus, W., 2022. Transactions on Machine Learning Research.

25. **Language Models are Few-Shot Learners**
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and others,, 2020. Advances in Neural Information Processing Systems, Vol 33.

26. **Multimodal Chain-of-Thought Reasoning in Language Models**
Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G. and Smola, A., 2023. arXiv preprint arXiv:2302.00923.

27. **Qwen2.5 Technical Report** [PDF]
Qwen,, :,, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z. and Qiu, Z., 2025.

28. **The Llama 3 Herd of Models**
Dubey, A., Jauhri, A., Pandey, A., Kadian, A. and others,, 2024. arXiv preprint arXiv:2407.21783.

29. **Llama 2: Open Foundation and Fine-Tuned Chat Models**
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. and others,, 2023. arXiv preprint arXiv:2307.09288.

30. **Mistral 7B**
Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L. and others,, 2023. arXiv preprint arXiv:2310.06825.

31. **Note on the sampling error of the difference between correlated proportions or percentages** [link]
Mcnemar, Q., 1947. Psychometrika, Vol 12, pp. 153-157.

32. **Statistical Power Analysis for the Behavioral Sciences**
Cohen, J., 1988. Lawrence Erlbaum Associates.

For attribution in academic contexts, please cite this work as

```
Anonymous, "The Feynman Trap: When Chain-of-Thought Reasoning Undermines Correct Intuitions in Language Models", TMLR Beyond PDF, 2026.
```

BibTeX citation

```
@inproceedings{anonymous2026thefeynmantrap,
  author = {Anonymous, },
  title = {The Feynman Trap: When Chain-of-Thought Reasoning Undermines Correct Intuitions in Language Models},
  booktitle = {TMLR Beyond PDF},
  year = {}
}
```