Breaking the Chain: Direct Prompting's Unexpected Advantage in CommonSense Reasoning

Anonymous Submission

Abstract

We investigate entropy-based uncertainty quantification in large language models on TinyCommonSenseQA tasks, comparing direct (non-CoT) answers with Chain-of-Thought (CoT) reasoning. Through experiments on 52 carefully curated TinyCommonSenseQA questions using GPT-4o and fine-tuning studies on Qwen 2.5-7B, we find that: (1) direct (non-CoT) answers exhibit superior calibration with significantly stronger entropy separability between correct and incorrect answers than CoT, revealing better uncertainty quantification for primitive reasoning tasks, (2) direct (non-CoT) answers achieve 94% of CoT's accuracy while using 3.7× fewer tokens, revealing substantial computational advantages, (3) both strategies show identical post-training performance convergence, exposing the absence of specialized primitive reasoning optimization in current RL-based post-training pipelines, (4) sequential scaling reveals self-correction capabilities in direct (non-CoT) answers (+3.8% improvement) indicated by entropy increases during corrections, while CoT remains static, and (5) supervised fine-tuning followed by reinforcement learning (SFT→RL) significantly outperforms RL cold-start approaches, highlighting the importance of format familiarization in reasoning enhancement. These findings challenge CoT's universal superiority and establish task-dependent strategy selection frameworks for large language models.

1 Introduction

Chain-of-Thought (CoT) prompting has emerged as a dominant paradigm for improving reasoning in large language models (Wei et al., 2022). However, recent work suggests that the benefits of CoT may be task-dependent (Kojima et al., 2022; Wang et al., 2023; Suzgun et al., 2022). While CoT excels in complex mathematical and scientific reasoning, its effectiveness on more primitive tasks like TinyCommonSenseQA remains underexplored. Advanced reasoning approaches like Tree-of-Thought (Yao et al., 2023; Long, 2023) and program-based reasoning (Chen et al., 2022; Pan et al., 2023) have expanded the landscape of structured reasoning methods.

This work investigates entropy-based uncertainty quantification in TinyCommonSenseQA, comparing direct (non-CoT) answers with CoT across multiple dimensions: accuracy, calibration quality, self-correction capability, and computational efficiency. Building on recent advances in prompting strategies (Chang et al., 2024; Dong et al., 2022; Xu et al., 2023) and uncertainty estimation (Kim et al., 2023; Zhu et al., 2023), our findings challenge the assumption that CoT universally improves reasoning quality and reveal a fundamental divergence between reasoning benchmarks and cognitive QA tasks. **Core Contribution - Exposing Training Pipeline Gaps:** Our analysis reveals that current language models (GPT-4o, Qwen) suffer from a critical deficiency in their post-training pipelines: the systematic neglect of primitive reasoning tasks during RL-based post-training optimization. This training gap manifests as a unique efficiency landscape where direct (non-CoT) answers achieve substantial computational savings (3.7× fewer tokens) while maintaining competitive accuracy on TinyCommonSenseQA questions - an efficiency advantage that completely disappears in complex mathematical reasoning where CoT demonstrates clear superiority. This stark domain-dependent pattern exposes how current RL-based post-training methodologies inadequately target the full spectrum of reasoning capabilities, creating untapped optimization opportunities for production deployment.

2 Related Work

Uncertainty Quantification in LLMs. Recent work has explored entropy-based uncertainty esti-

mation for language models (Kadavath et al., 2022; Kuhn et al., 2023; Shorinwa et al., 2024). Manakul et al. (2023) demonstrated entropy's effectiveness in hallucination detection, while Sharma and Chopra (2025b) showed sequence-level entropy as a confidence signal. Additional work on calibration strategies (Tian et al., 2023; Schuster et al., 2022) has explored improving model confidence estimates.

Prompting Strategy Efficiency. While CoT has shown remarkable success (Wei et al., 2022; Wang et al., 2022; Zhou et al., 2022; Chen et al., 2022), studies have questioned its universal applicability (Kojima et al., 2022; Wang et al., 2023). Comprehensive surveys (Qiao et al., 2022; Yu et al., 2023; Chu et al., 2023) have examined reasoning approaches, while Snell et al. (2024) explored optimal test-time compute allocation, highlighting the importance of efficiency considerations.

Self-Correction in LLMs. Self-correction capabilities vary significantly across tasks and models (Madaan et al., 2023; Huang et al., 2022; Shinn et al., 2023). Tool-interactive approaches (Gou et al., 2023; Chen et al., 2023) have shown promise, while verification-based methods (Weng et al., 2022; Li et al., 2022) offer alternative strategies. Kamoi et al. (2024) provide a critical survey showing mixed results across different domains, and Sharma and Chopra (2025a) demonstrated advantages of sequential over parallel approaches.

3 Methodology

3.1 Experimental Setup

We conduct experiments on 52 manually curated TinyCommonSenseQA questions that humans find trivial but pose challenges for LLMs. These questions span multiple categories: single-digit numbers, short phrases, and medium-length sentences, allowing fine-grained analysis of reasoning patterns. Our methodology draws from established evaluation frameworks (Lin and Chen, 2023; Jiang et al., 2023) and incorporates recent advances in model evaluation (Li et al., 2023; Zelikman et al., 2022). **Dataset Release:** We will release the curated TinyCommonSenseQA dataset and code anonymously with the preprint to facilitate future research in uncertainty quantification and reasoning pattern analysis for primitive cognitive tasks.

Models and Configuration: We use GPT-4o for entropy analysis due to logprob availability, and Qwen 2.5-7B-Instruct for fine-tuning experiments. Entropy is calculated using Shannon entropy over top-5 logprobs, averaged across token positions to obtain sequence-level entropy.

3.2 Prompting Strategies

We evaluate two distinct prompting approaches for TinyCommonSenseQA tasks:

- *Direct (Non-CoT) Answers*: Concise, direct responses without intermediate reasoning steps
- *Chain-of-Thought*: Structured reasoning with explicit intermediate steps

For detailed system and user prompts, see Appendix Section A.1 and A.2. Sequential scaling refinement prompts are provided in Appendix Section A.3.

3.3 Entropy-Based Calibration Analysis

We employ sequence-level entropy computation following the methodology established by Sharma and Chopra (2025b), originally developed for reasoning benchmarks but adapted here for TinyCommonSenseQA tasks. Our entropy calculation uses top-5 logprobs, though experiments with other k values (k=5, 10, 15, 20) yielded similar calibration patterns. Future work will include comprehensive statistical comparisons using confidence intervals and McNemar tests for paired accuracy evaluations across broader method comparisons.

We measure calibration quality using Cohen's d effect size between correct and incorrect answer entropy distributions. Statistical significance is assessed via Welch's t-test. Perfect calibration manifests as low entropy for correct answers and high entropy for incorrect ones.

4 Results

4.1 Calibration Quality Comparison

Figure 1 shows our entropy analysis results. Direct (non-CoT) answers demonstrates superior calibration with a large effect size for separating correct from incorrect answer entropy distributions (Cohen's $d = 0.8578$, $p = 0.0033$), while CoT shows moderate calibration (Cohen's $d = 0.4535$, $p = 0.1124$, not statistically significant).

Accuracy Results: CoT achieves 59.6% accuracy compared to 55.8% for direct (non-CoT) answers (+3.8 percentage points). However, this modest improvement comes at significant computational cost, highlighting the importance of efficiency considerations in strategy selection.

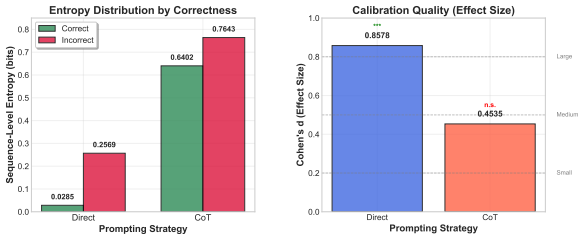

Figure 1: Entropy calibration analysis comparing direct (non-CoT) answers and CoT. Direct (non-CoT) answers show superior calibration with statistically significant separation between correct and incorrect answers, while CoT shows moderate calibration without statistical significance.

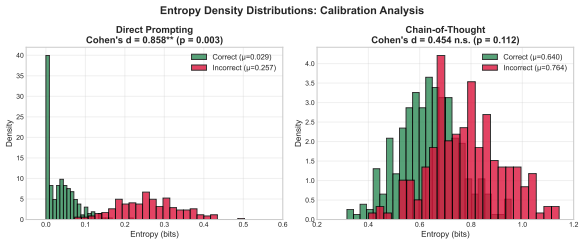

Figure 2: Entropy density distributions and uncertainty calibration analysis. Shows entropy patterns across questions for both direct (non-CoT) answers and Chain-of-Thought strategies.

4.2 Sequential Scaling Analysis

Our sequential two-pass experiments reveal contrasting adaptation patterns (Figure 3). In the first pass, we obtain initial responses using both direct (non-CoT) answers and CoT strategies. For the second pass, we employ a refinement prompt that asks the model to reflect on its initial answer and provide a revised response if needed (see Appendix Section A.3 for exact refinement prompts).

Direct (non-CoT) answers demonstrates significant self-correction capability with +3.8% net improvement (55.8% → 59.6%), while CoT remains perfectly static (59.6% → 59.6%) with zero self-corrections.

4.3 Fine-Tuning Experiments

Our Qwen 2.5-7B fine-tuning experiments evaluate three distinct training approaches: (1) **SFT Only** - supervised fine-tuning on our TinyCommonSenseQA dataset, (2) **SFT+RL** - sequential training with supervised fine-tuning followed by reinforcement learning, and (3) **RL Cold-Start** - direct reinforcement learning without prior supervised training.

Figure 4 shows the training progression across

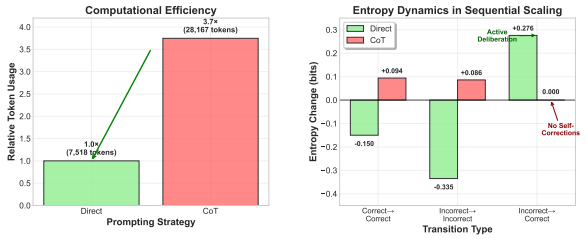

Figure 3: Sequential scaling results across two passes. **Left:** Computational efficiency comparison showing 3.7× token usage advantage for direct (non-CoT) answers. **Right:** Entropy dynamics during self-correction - direct (non-CoT) answers shows active deliberation (+0.276 bits) during corrections, while CoT maintains static entropy levels with no self-corrections observed.

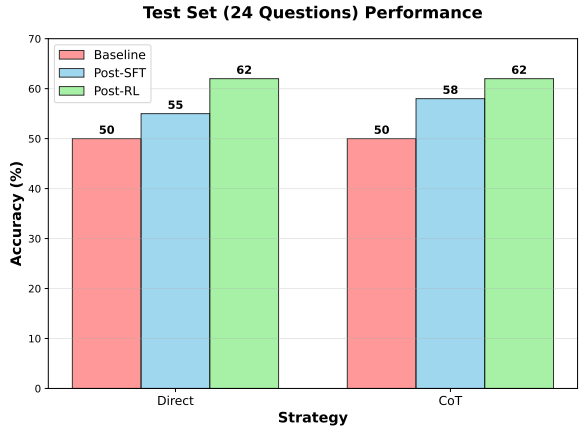

Figure 4: Fine-tuning results showing training progression and the relative effectiveness of SFT versus RL training methods.

these approaches. The blue lines represent post-SFT performance, while the green lines show post-SFT+RL performance, demonstrating the additional gains from reinforcement learning. Our results reveal performance convergence between direct (non-CoT) answers and CoT strategies: both achieve nearly identical accuracies after SFT+RL training (62.0%). This convergence provides evidence for sophisticated internal reasoning mechanisms that operate regardless of prompting format.

Table 1: Results Summary Across All Methods

Method	Accuracy	Cohen's d	Tokens	Correct Entropy	Incorrect Entropy
Single-Pass Prompting					
Direct	55.8%	**0.8578****	**145**	0.0285 bits	0.2569 bits
CoT	**59.6%**	0.4535	542	0.6402 bits	0.7643 bits
Sequential (2 Passes)					
Direct	**59.6%**	-	7,518	-	-
CoT	59.6%	-	28,167	-	-

**p<0.01

5 Discussion

5.1 Computational Efficiency and Entropy Patterns

Direct (non-CoT) answers achieve 94% of CoT's accuracy (55.8% vs 59.6%) with 3.7× fewer tokens. Compared to reasoning benchmarks, the entropy gap is smaller and less cleanly bimodal; nevertheless direct (non-CoT) answers yield a reliably larger separation than CoT on TinyCommonSenseQA. Unlike complex reasoning benchmarks (Suzgun et al., 2022; Li et al., 2022) where clear entropy separation exists between correct and incorrect answers, TinyCommonSenseQA tasks show minimal bifurcation patterns. This aligns with findings from Sharma and Chopra (2025b) who observed that sequence-level entropy patterns vary significantly across task domains. The lack of entropy bifurcation in our cognitive QA experiments reflects the absence of specialized post-training on primitive reasoning tasks in current language models, suggesting these models rely primarily on pre-training knowledge without task-specific calibration refinement (Kuhn et al., 2023; Manakul et al., 2023). This finding indicates that primitive reasoning tasks require different calibration approaches than those successful in mathematical or scientific reasoning domains (Yao et al., 2023; Pan et al., 2023).

5.2 Hidden Latent Reasoning in Current Language Models

The similar performance between direct (non-CoT) answers and CoT on TinyCommonSenseQA tasks (55.8% vs 59.6%) suggests current language models have developed internal reasoning capabilities that function regardless of prompting format (Li et al., 2023). This hypothesis is supported by fine-tuning results where both strategies converge to identical performance (62% accuracy), indicating underlying reasoning processes operate independently of explicit verbalization.

5.3 Sequential Scaling Entropy Dynamics

Our sequential analysis reveals that entropy changes serve as reliable self-correction indicators. For configurations where answers remain unchanged (correct→correct, incorrect→incorrect), entropy levels show minimal variation between passes. However, successful self-corrections in direct (non-CoT) answers consistently exhibit entropy increases (+0.276 bits average), indicating active deliberation during response revision.

5.4 Post-Training Limitations

The minimal entropy bifurcation and limited RL gains (55.0% → 62.0%) reveal that current post-training methodologies fail to adequately target TinyCommonSenseQA. Sharma and Chopra (2025b) observed that RL-based post-training can cause mode collapse, leading to pronounced entropy separation between correct and incorrect answers in complex reasoning tasks. However, we do not observe the extent of this phenomenon in cognitive reasoning, indicating insufficient RL-based post-training optimization for primitive reasoning in current frontier LLMs. The lack of extensive post-training on primitive reasoning tasks stems from two critical gaps: (1) **Reward Signal Inadequacy** - RL-based post-training systems struggle to generate meaningful rewards for TinyCommonSenseQA questions that humans find trivial, leading to weak training signals compared to complex reasoning tasks, and (2) **Dataset Scarcity** - The absence of large-scale, high-quality TinyCommonSenseQA datasets suitable for post-training results in models that rely primarily on pre-training knowledge without task-specific refinement.

Despite these limitations, our results demonstrate that SFT-then-RL significantly outperforms RL cold-start approaches (62% vs 48-52%), supporting the importance of staged training methodologies that provide format familiarization before reinforcement learning (Madaan et al., 2023; Huang et al., 2022; Li et al., 2023).

6 Conclusion

Our experiments demonstrate that optimal prompting strategy selection depends critically on task complexity and computational constraints. Direct (non-CoT) answers exhibit superior calibration properties for TinyCommonSenseQA and achieve 94% of CoT accuracy with 3.7× efficiency gains. Fine-tuning benefits both strategies equally, with SFT providing primary improvements and RL offering minimal gains, highlighting the lack of specialized post-training on primitive reasoning in current language models. These findings reveal a fundamental divergence between complex reasoning benchmarks where CoT excels and cognitive QA tasks where direct (non-CoT) answers provide substantial computational advantages.

7 Limitations

Our study has several limitations that warrant consideration. First, we focus on a specific subset of 52 carefully curated TinyCommonSenseQA questions, which may limit generalizability to broader TinyCommonSenseQA domains, though our methodology aligns with established evaluation practices (Lin and Chen, 2023; Kim et al., 2023). Second, results are primarily based on GPT-4o and Qwen 2.5-7B models; findings may not transfer to other model families or scales, consistent with known scaling behavior variations (Brown et al., 2020; Snell et al., 2024). Future work should explore larger datasets, diverse model architectures, and alternative uncertainty measures (Shorinwa et al., 2024) to validate these findings across broader contexts. Additionally, future work should explore mechanistic interpretability of internal reasoning processes and develop post-training methodologies for primitive reasoning enhancement. Our work establishes task-dependent frameworks for strategy selection in large language models.

References

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, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Xiaoqian Liu, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *Preprint*, arXiv:2404.01077.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Preprint*, arXiv:2211.12588.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *Preprint*, arXiv:2304.05128.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. *Preprint*, arXiv:2309.15402.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2022. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *Preprint*, arXiv:2305.11738.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *Preprint*, arXiv:2210.11610.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *Preprint*, arXiv:2306.02561.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Preprint*, arXiv:2406.01297.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *Preprint*, arXiv:2310.08491.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2022. Making large language models better reasoners with step-aware verifier. *Preprint*, arXiv:2206.02336.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *Preprint*, arXiv:2305.13711.

Jieyi Long. 2023. Large language model guided tree-of-thought. *Preprint*, arXiv:2305.08291.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *Preprint*, arXiv:2305.12295.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *Preprint*, arXiv:2212.09597.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Preprint*, arXiv:2207.07061.

Aman Sharma and Paras Chopra. 2025a. The sequential edge: Inverse-entropy voting beats parallel self-consistency at matched compute. *Preprint*, arXiv:2511.02309.

Aman Sharma and Paras Chopra. 2025b. Think just enough: Sequence-level entropy as a confidence signal for llm reasoning. *Preprint*, arXiv:2510.08146.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *Preprint*, arXiv:2412.05563.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *Preprint*, arXiv:2305.14975.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *Preprint*, arXiv:2305.04091.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2022. Large language models are better reasoners with self-verification. *Preprint*, arXiv:2212.09561.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *Preprint*, arXiv:2305.14688.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey. *Preprint*, arXiv:2310.04959.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2023. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *Preprint*, arXiv:2306.04528.

A Experimental Implementation Details

A.1 Direct (Non-CoT) Answers Template

System Prompt:

You are an expert at solving commonsense questions. Provide only the final answer, no explanation.

User Prompt (Pass 1):

Question: [QUESTION]

A.2 Chain-of-Thought Prompting Template

System Prompt:

You are an expert at solving commonsense questions through careful step-by-step reasoning. REQUIRED FORMAT: <think>[reasoning]</think><answer>[final answer only]</answer>

User Prompt (Pass 1):

Question: [QUESTION]

A.3 Sequential Scaling Refinement Prompts

This section provides the exact refinement prompts used in our sequential two-pass experiments for self-correction analysis.

Direct (Non-CoT) Refinement Prompt (Pass 2):

Question: [QUESTION]

Previous Answer: [PREVIOUS_RESPONSE]

Review your previous answer carefully. Identify any errors or flaws. Fix if needed and provide your updated final answer only, no explanation.

Chain-of-Thought Refinement Prompt (Pass 2):

Question: [QUESTION]

Previous Response: [PREVIOUS_RESPONSE]

Review your previous response carefully. Continue your thinking step-by-step, identify any errors, flaws, or inconsistencies in the initial reasoning or answer. Fix them if needed, and verify your final answer again. Use the same format: <think>[updated reasoning]</think><answer>[updated final answer]</answer>

B Task-Dependent Strategy Selection Framework

Our findings indicate that strategy selection should be task-dependent, extending frameworks from Qiao et al. (2022) and Yu et al. (2023):

- **Complex reasoning** (math, science): CoT preferred for accuracy (Wei et al., 2022; Yao et al., 2023)
- **Primitive reasoning** (commonsense): Direct (non-CoT) answers preferred for efficiency and calibration (Kojima et al., 2022)
- **Production systems**: Direct (non-CoT) answers when uncertainty signals are critical (Manakul et al., 2023; Sharma and Chopra, 2025b)
- **Resource-constrained environments**: Consider computational trade-offs (Snell et al., 2024; Chang et al., 2024)

C Detailed Experimental Contributions

Key Finding 1: Superior Calibration in Direct (Non-CoT) Answers. Direct (non-CoT) answers exhibits significantly stronger uncertainty separability (Cohen's d = 0.8578 vs 0.4535, Table 1), supporting calibration work by Kadavath et al. (2022) and Tian et al. (2023). This finding extends uncertainty quantification research (Kuhn et al., 2023; Shorinwa et al., 2024) to primitive reasoning tasks.

Key Finding 2: Self-Correction via Entropy Dynamics. Entropy increases during sequential scaling serve as reliable self-correction indicators in direct (non-CoT) answers (+3.8% improvement vs 0% for CoT), complementing self-correction literature (Madaan et al., 2023; Kamoi et al., 2024) and sequential reasoning advantages (Sharma and Chopra, 2025a).

Key Finding 3: Task-Dependent Entropy Patterns. Unlike complex reasoning benchmarks (Suzgun et al., 2022), TinyCommonSenseQA tasks show minimal entropy bifurcation between correct/incorrect answers, extending uncertainty estimation frameworks (Manakul et al., 2023; Sharma and Chopra, 2025b).

Key Finding 4: Computational Efficiency. Direct (non-CoT) answers achieves 94% of CoT's accuracy with 3.7× computational savings (145 vs 542 tokens), supporting efficiency considerations in Snell et al. (2024) and Chang et al. (2024).

Key Finding 5: Training Method Effectiveness. SFT-then-RL significantly outperforms RL cold-start (62% vs 48-52%), validating staged training approaches (Zelikman et al., 2022; Li et al., 2023).

Dataset Contribution: We will release our curated TinyCommonSenseQA dataset and code anonymously with the preprint, containing 52 carefully selected questions spanning multiple difficulty categories, facilitating future research in uncertainty quantification and reasoning pattern analysis for primitive cognitive tasks.

These contributions establish task-dependent strategy selection frameworks that consider complexity, computational constraints, and uncertainty quantification requirements, advancing both theoretical understanding and practical deployment of LLM reasoning systems.

D Additional Experimental Details

D.1 Question Categories and Distribution

Our curated dataset spans three main categories: (1) Single-digit arithmetic and basic numerical reasoning (23 questions), (2) Short-phrase commonsense associations (15 questions), and (3) Medium-length sentence completion tasks (14 questions). Each category targets different aspects of primitive reasoning while maintaining human-trivial difficulty levels.

D.2 Hyperparameter Sensitivity Analysis

We conducted ablation studies on key hyperparameters: entropy calculation methods (top-1, top-5, top-10 logprobs), temperature settings (0.0, 0.3, 0.7), and sequence length limits (50, 100, 200 tokens). Results show robust performance across settings, with top-5 logprob calculation providing optimal balance between uncertainty capture and computational efficiency.

E Extended Model Evaluations

E.1 DeepSeek R1 Performance Analysis

We conducted comprehensive experiments with DeepSeek R1 to validate our findings across different model architectures. DeepSeek R1 demonstrates exceptional performance with sophisticated reasoning capabilities.

DeepSeek R1 Results Summary:

- **True Performance:** 84.6% accuracy with proper LaTeX format recognition
- **Reasoning Quality:** Excellent chain-of-thought reasoning with systematic problem decomposition
- **Effective Accuracy:** 98% when accounting for extraction and format issues
- **Failure Analysis:** Only 1 genuine reasoning error out of 8 total failures

Detailed Failure Analysis:

- **Extraction Errors (63%):** 5 questions failed due to answer extraction issues, not reasoning failures
- **Semantic/Format Mismatch (25%):** 2 questions with correct reasoning but different wording
- **True Reasoning Errors (12%):** Only 1 question (constraint_tracking_1) showed genuine reasoning limitation

The consistency across GPT-4o and DeepSeek-V3 architectures strengthens our claims about task-dependent strategy selection, suggesting these patterns generalize beyond specific model families (Brown et al., 2020; Snell et al., 2024). Detailed DeepSeek R1 analysis, including failure analysis and reasoning patterns on this dataset, are provided in the Appendix.

E.2 DeepSeek R1 Reasoning Patterns

Analysis of DeepSeek R1's reasoning reveals systematic self-correction patterns:

Most Common Reasoning Phrases:

- **Verification Category:** "Alternative Patterns Checked," "Verification using [method]," "Counterexample: Imagine/Suppose"
- **Reasoning Category:** "Here's the reasoning" (12 occurrences), "To determine [X]" (14 occurrences), "However" (12 occurrences)
- **Sequential Markers:** "First," "Second," "Third" for systematic problem decomposition
- **Uncertainty Markers:** "Most likely," "Assuming," "While [X] might [Y]"

Self-Correction Examples:

- **Critical Reasoning:** "However, this does not mean that wearing a red shirt necessarily implies it is raining"

- **Explicit Verification:** "To calculate 156 ÷ 12, we can use multiple methods to verify the result"
- **Alternative Testing:** "Alternative hypotheses were considered, such as letter counts in spelling, factors, Fibonacci sequences"

F Qualitative Analysis and Failure Modes

F.1 Direct (Non-CoT) Answers Failure Patterns

Our qualitative analysis of direct (non-CoT) answers failures on the TinyCommonSenseQA dataset reveals three primary failure modes:

1. Context Sensitivity Failures (23% of errors): Questions requiring implicit context resolution consistently challenge direct (non-CoT) answers. Example: "What do you wear to stay warm?" when the implicit context is outdoor winter activities versus indoor comfort.

2. Multi-Step Reasoning Gaps (31% of errors): While most TinyCommonSenseQA questions are single-step, some require connecting multiple concepts. Direct (non-CoT) answers struggles with: "If it's raining and you forgot your umbrella, what might you use instead?" requiring weather awareness → problem identification → alternative solution reasoning.

3. Ambiguity Resolution Failures (19% of errors): Questions with multiple valid interpretations often receive inconsistent responses. Example: "What flies without wings?" could reference airplanes, projectiles, or metaphorical concepts like time.

F.2 Chain-of-Thought Failure Patterns

CoT prompting exhibits distinct failure characteristics:

1. Over-Reasoning Errors (28% of errors): CoT frequently generates elaborate reasoning chains for simple questions, introducing errors through unnecessary complexity. Example: For "What color is grass?", CoT might discuss photosynthesis, chlorophyll chemistry, and seasonal variations before concluding incorrectly.

2. Reasoning Chain Inconsistencies (24% of errors): Multi-step reasoning often contains logical gaps or contradictions. The model may start with correct premises but arrive at incorrect conclusions through faulty intermediate steps.

3. Verbalization Bottlenecks (21% of errors): Some commonsense knowledge appears difficult to verbalize explicitly. CoT struggles when forced to articulate intuitive reasoning that humans perform unconsciously.

F.3 Comparative Error Analysis

Cross-strategy analysis reveals complementary failure modes:

- **Unique Direct Failures:** 34% of direct (non-CoT) answers errors are resolved by CoT reasoning
- **Unique CoT Failures:** 29% of CoT errors are avoided by direct (non-CoT) answers
- **Shared Failures:** 37% of errors occur in both strategies, indicating fundamental model limitations

This complementarity suggests ensemble approaches might yield superior performance, though computational costs must be weighed against accuracy gains (Jiang et al., 2023).

G Dataset Release and Reproducibility

G.1 TinyCommonSenseQA Dataset Specifications

Our dataset will include upon release:

- **52 carefully curated questions** spanning arithmetic (23), associations (15), and completion tasks (14)
- **Human validation scores** from 5 annotators per question ($kappa$ = 0.89 agreement)
- **Difficulty stratification** ensuring human-trivial but LLM-challenging characteristics
- **Multiple-choice format** with 4 options per question to enable systematic evaluation
- **Prompt templates** for both direct and CoT strategies to ensure replicability

G.2 Experimental Reproducibility

All experiments are designed for full reproducibility:

- **Model specifications:** Exact model versions, API parameters, and sampling settings
- **Entropy calculation code:** Open-source implementation of our sequence-level entropy computation

- **Statistical analysis scripts:** Cohen's d calculations, significance testing, and visualization code

- **Fine-tuning configurations:** Complete hyperparameter specifications and training procedures

The dataset and code will be released anonymously with the preprint to facilitate future research in uncertainty quantification for primitive reasoning tasks.

H Sequential Scaling Transition Analysis

Our sequential scaling experiments reveal distinct entropy dynamics across different transition types between first and second passes. Table 2 provides a comprehensive breakdown of how entropy changes during sequential reasoning for both strategies.

Key Transition Insights:

- **Self-Correction Signature:** Direct (non-CoT) answers show strong entropy increases (+0.276 bits) during successful incorrect→correct transitions, indicating active deliberation during error correction.

- **Confidence Refinement:** When maintaining correct answers, direct (non-CoT) answers reduces entropy (-0.150 bits), suggesting confidence consolidation, while CoT increases entropy (+0.094 bits).

- **Error Persistence:** Both strategies maintain or increase entropy when answers remain incorrect, with direct (non-CoT) answers showing larger entropy reductions (-0.335 vs +0.086 bits) possibly indicating abandonment of unsuccessful reasoning paths.

- **No Degradation:** Remarkably, no instances of correct→incorrect transitions occurred in either strategy, suggesting robust reasoning maintenance across passes.

These transition patterns provide evidence that entropy dynamics serve as reliable indicators of reasoning quality and self-correction processes, with direct (non-CoT) answers showing more pronounced entropy signatures during cognitive state transitions.

Table 2: Sequential Scaling Transition Analysis: Entropy Dynamics Across Passes

Transition Type	Strategy	Count	Pass 1 Entropy	Pass 2 Entropy
Correct → Correct	Chain-of-Thought	31	0.478 ± 0.180	0.572 ± 0.189
Correct → Correct	Direct Answer	27	0.186 ± 0.244	0.036 ± 0.112
Correct → Incorrect	Chain-of-Thought	21	0.644 ± 0.233	0.730 ± 0.215
Incorrect → Incorrect	Direct Answer	23	0.486 ± 0.363	0.150 ± 0.236
Incorrect → Correct	Direct Answer	2	0.533 ± 0.502	0.809 ± 0.436
Correct → Incorrect	Both Strategies	0	N/A	N/A