Implicit Chain-of-Thought for Abstractive Text Summarization: A Fast, Accurate Alternative to Explicit CoT

Anonymous ACL submission

Abstract

Explicit Chain-of-Thought (CoT) improves LLM reasoning but is slow because models must generate long scratchpad text; No-CoT is fast but less accurate. Recent work shows that *Implicit CoT*, internalizing reasoning in hidden states and emitting only the final answer, can be nearly as accurate as Explicit CoT while running at No-CoT speeds. Prior studies focus on arithmetic and math word problems; to our knowledge, *no work has evaluated Implicit CoT for text summarization*. We present the first study of Implicit CoT for *abstractive* dialogue summarization on the SAMSum dataset using stepwise internalization, comparing No CoT, Explicit CoT, and Implicit CoT trained on GPT-Neo 1.3B. Our results demonstrate that Implicit CoT achieves 98.4% of Explicit CoT's ROUGE-1 performance (0.1929 vs 0.1960) while training 3.6 times faster (3.1 hours vs 11.4 hours), maintaining inference efficiency comparable to No-CoT, and thus bridging the gap between quality and computational efficiency.

1 Introduction

Chain-of-Thought (CoT) prompting improves multi-step problem solving by externalizing intermediate reasoning, but it increases latency and token cost (Wei et al., 2022; Chu et al., 2024). Strong explicit variants include self-consistency, decomposition, search over thoughts, and retrieval-interleaved reasoning, which further boost accuracy while often producing even longer traces (Xia et al., 2025; Besta et al., 2024). Recent results suggest that models can internalize useful computations when trained with stepwise removal of visible rationales or with soft latent thoughts, recovering much of the benefit of explicit reasoning while reducing tokens (Deng et al., 2025; Xu et al., 2025).

This work distinguishes three regimes: *No-CoT* predicts the summary directly; *Explicit CoT* emits a visible scratchpad before the final summary; *Implicit CoT* targets internal computation with minimal surface text. A key risk is shortcut behavior under weak supervision (Lin et al., 2025), which we address with stable, fixed-pattern planning traces and curriculum-based removal schedules.

We trained and evaluated these approaches for dialogue summarization on SAMSum using GPT Neo 1.3B. Results confirm that Explicit CoT significantly improves quality over No-CoT (61.2% relative improvement in ROUGE-1). Critically, our stepwise internalization approach achieves Implicit CoT performance within 1.6% of Explicit CoT while reducing training time by 72%, demonstrating the viability of implicit reasoning for summarization tasks.

Contributions. Our main contributions are:

- We present the first study of Implicit Chain-of-Thought for abstractive dialogue summarization.
- We show that stepwise internalization achieves 98.4% of Explicit CoT performance while reducing training time by 72%.
- We demonstrate that Implicit CoT bridges the quality–efficiency gap, achieving near-Explicit quality with near-No-CoT inference cost.

2 Background

Text summarization condenses documents into concise summaries while preserving meaning. We focus on abstractive summarization, which generates paraphrastic summaries using novel wording.

How to train Implicit CoT (two families). Training Implicit CoT has two main approaches: **ICoT-KD** (knowledge distillation) transfers behaviors from an explicit CoT teacher to a student performing implicit reasoning, while **ICoT-SI** (stepwise internalization) fine-tunes an Explicit-CoT

model while gradually removing CoT tokens, with optimizer resets and mild randomization to stabilize the curriculum (Deng et al., 2025).

In stepwise internalization, the expected mechanism is that explicit plans improve content selection and organization; gradual internalization compresses these computations into hidden states, retaining planning benefits without emitting long chains. To stabilize this curriculum, the optimizer is reset between removal stages and mild randomization is applied to the removal schedule. In practice, stepwise internalization yields strong speed-accuracy trade-offs for implicit reasoning while minimizing emitted tokens (Deng et al., 2025; Xu et al., 2025).

3 Related Work

Wei et al. (2022) introduced Chain-of-Thought prompting, showing that explicit intermediate reasoning improves multi-step problem solving. Subsequent work developed stronger variants including self-consistency (Xia et al., 2025), Least-to-Most decomposition (Xia et al., 2025), Tree-of-Thoughts search (Besta et al., 2024), and graph-structured reasoning (Besta et al., 2024; Han et al., 2025). These approaches are foundational to modern LLM-based autonomous agents (Wang et al., 2024; Xi et al., 2025; Sumers et al., 2024), which increasingly employ planning and tool use (Huang et al., 2024; Qin et al., 2024).

For efficiency, implicit approaches remove visible rationales and internalize computation; stepwise internalization and soft latent thoughts retain much of CoT's benefit with far fewer output tokens (Deng et al., 2025; Xu et al., 2025); latent chains can sometimes be surfaced without prompts (Wang and Zhou, 2024), though weak supervision risks shortcuts (Lin et al., 2025). Reflexive agents further enhance reasoning through self-correction mechanisms (Shinn et al., 2023; Zhang et al., 2024).

Most prior evaluations focus on arithmetic or math word problems; we study implicit CoT for *abstractive summarization*, assessing faithfulness and efficiency side by side.

4 Methodology

4.1 Problem Statement

Given a source document, produce an abstractive summary that is concise and faithful while minimizing inference cost. The goal is to determine whether stepwise internalization can retain the planning benefits of explicit reasoning without emitting scratchpad tokens, thereby improving the accuracy-cost trade-off (Chu et al., 2024; Deng et al., 2025).

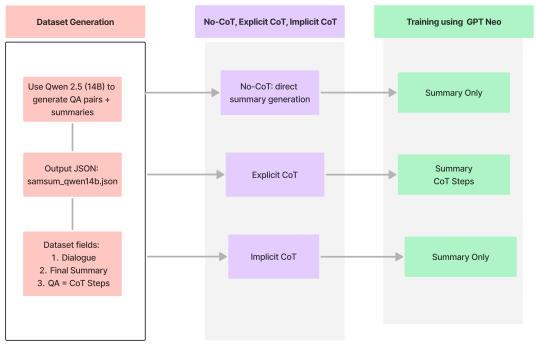

Figure 1: Workflow: Dataset generation using Qwen 2.5 (14B) creates structured training data, followed by training GPT Neo 1.3B under three reasoning regimes.

4.2 Workflow and Data

Figure 1 illustrates our approach. We use Qwen 2.5 (14B) to generate structured JSON files from the SAMSum dialogue dataset (14,700 samples with human-written summaries), containing dialogue-grounded QA pairs and CoT reasoning steps. We use standard 80%/10%/10% train/validation/test splits.

4.3 Model and Training

GPT Neo 1.3B serves as our primary model, chosen for its balance between capacity and computational efficiency. We employ LoRA fine-tuning ($r = 16$) with AdamW optimization (learning rate 3×10^{-5}, batch size 4 with gradient accumulation over 8 steps, gradient clipping at 1.0). All hyperparameters remain identical across regimes to isolate the effect of reasoning strategy.

4.4 Training Regimes

No CoT maps dialogue directly to summary. *Explicit CoT* uses a structured scratchpad (*plan* $\rightarrow$ *key points* $\rightarrow$ *summary*), generating the reasoning trace before the final summary (Wei et al., 2022). *Implicit CoT* adopts stepwise internalization (Deng et al., 2025): training begins as Explicit CoT, then progressively removes initial CoT tokens across fine-tuning stages, resetting the optimizer at each stage with mild stochastic offset to smooth the objective. At the final stage, the model emits only the summary.

The expected mechanism is that explicit plans improve content selection and organization; gradual internalization compresses these computations into hidden states, retaining planning benefits without emitting long chains. Comparisons include No-CoT, Explicit CoT, and the proposed Implicit CoT, with evaluation spanning automatic metrics (ROUGE, BERTScore) and training efficiency.

4.5 Stepwise Internalization Implementation

Our implementation of stepwise internalization follows the curriculum learning approach proposed by Deng et al., with specific design choices tailored to dialogue summarization:

Removal schedule. We train for 8 epochs with a removal rate of $\Delta = 8$ CoT tokens per epoch. The structured scratchpad contains approximately 64 tokens on average (plan + key points), so this schedule achieves near-complete internalization by epoch 8. CoT tokens are removed from left to right (beginning of the reasoning chain), following the principle that early planning steps should be internalized first while later verification steps may remain visible longer.

Optimizer reset. To prevent catastrophic forgetting when the input distribution shifts (due to removed tokens), we reset the AdamW optimizer state at the beginning of each epoch. This allows the model to re-adapt its learning dynamics to the modified objective without carrying over momentum from the previous token configuration.

Removal smoothing. To avoid abrupt objective changes, we apply stochastic smoothing with a small probability of adding a random offset to the scheduled removal count. This mild randomness helps the model generalize across slightly different internalization stages.

4.6 Evaluation

We use ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum for n-gram overlap, plus BERTScore F1 for semantic similarity. We report training duration, throughput (samples/sec), and inference characteristics (token count and latency).

5 Results

Table 1 shows that Implicit CoT achieves ROUGE-1 of 0.1929, representing 98.4% of Explicit CoT's performance (0.1960), with only a 0.0031 gap. This demonstrates successful internalization where the model retains planning benefits while eliminating scratchpad generation. Implicit CoT outperforms No-CoT by 61.2% relative improvement in ROUGE-1, with similar gains across ROUGE-2 (40.8%), ROUGE-L (62.3%), and BERTScore (1.2%), validating that internalized reasoning substantially improves summarization quality.

Implicit CoT trains in 3h 8m versus 11h 27m for Explicit CoT, a 72% reduction while maintaining 98.4% of the quality. Both No-CoT and Implicit CoT emit only the final summary, while Explicit CoT generates the full reasoning chain, making Implicit CoT approximately $4\times$ faster at inference. This positions Implicit CoT optimally: near-Explicit quality with near-No-CoT efficiency.

The additional Δ columns explicitly report the absolute performance differences between Implicit CoT and the No-CoT and Explicit CoT baselines, making the relative gains and remaining gaps directly observable.

6 Discussion

6.1 Why Implicit CoT Succeeds for Summarization

The success of Implicit CoT stems from three factors. First, summarization planning is compressible; content selection and organization can be effectively internalized into distributed representations, unlike arithmetic where each step must be explicit. Second, stepwise internalization provides stable supervision through gradual token removal with optimizer resets and smoothing, teaching the model to fill in missing reasoning rather than adopt shortcuts. Third, SAMSum dialogues follow recognizable patterns (greetings, discussions, conclusions), enabling reliable internal heuristics.

6.2 Remaining Quality Gap

The 1.6% gap between Implicit and Explicit CoT suggests partial internalization. Contributing factors include: (1) the 8-epoch curriculum length may benefit from refinement, (2) removing 8 tokens per epoch is relatively aggressive, and (3) GPT-Neo 1.3B has limited hidden capacity for complex reasoning. Despite this, the quality-cost trade-off strongly favors Implicit CoT for deployment.

6.3 Training and Inference Efficiency

The $3.6\times$ training speedup stems from progressive simplification, as CoT tokens are removed, generation becomes easier, enabling faster convergence.

Table 1: Experimental results comparing No-CoT, Implicit CoT, and Explicit CoT on SAMSum dialogue summarization. All methods use GPT-Neo 1.3B with identical hyperparameters.

Metric	**No CoT**	**Implicit CoT**	**Explicit CoT**	ΔImplicit–No CoT	ΔImplicit–Explicit
Quality Metrics					
ROUGE-1	0.1197	0.1929	**0.1960**	+0.0732	-0.0031
ROUGE-2	0.0471	0.0663	**0.0824**	+0.0192	-0.0161
ROUGE-L	0.0954	0.1548	**0.1723**	+0.0594	-0.0175
ROUGE-Lsum	0.0958	**0.1632**	0.1583	+0.0674	+0.0049
BERTScore F1	0.8352	0.8453	**0.8532**	+0.0101	-0.0079
Training Efficiency					
Training Duration	1h 27m	3h 08m	11h 27m	–	–
Throughput (samples/s)	**22.49**	11.45	2.85	–	–
Inference Characteristics					
Output Tokens	Low	Low	High	–	–
Inference Speed	**Fastest**	Fast (4×)	Slowest	–	–

Reduced output length also improves throughput directly. Implicit CoT trains slower than No-CoT (3.1 vs 1.4 hours) because early epochs still generate partial CoT, but this overhead is the price of learning internalized reasoning that pays dividends in quality.

By emitting only summaries while maintaining near-Explicit quality, Implicit CoT reduces per-request latency, token usage, and API costs by approximately 4× compared to Explicit CoT. This enables reasoning-enhanced summarization for latency-sensitive applications like real-time chat summarization, news aggregation, and conversational AI where inference cost dominates training cost.

7 Limitations

This study evaluates only GPT-Neo 1.3B due to GPU availability. While sufficient to demonstrate the Implicit CoT concept, larger models (2.7B+) may achieve better internalization and close the remaining quality gap. Cross-model validation on LLaMA and Qwen architectures would strengthen generalizability but was not feasible within available compute resources.

8 Conclusion

This work presents the first evaluation of Implicit Chain-of-Thought for abstractive summarization via stepwise internalization. Training GPT-Neo 1.3B on SAMSum, we demonstrate that Implicit CoT achieves 98.4% of Explicit CoT's ROUGE-1 performance while reducing training time by 72% and maintaining inference efficiency comparable to No-CoT. This bridges the quality-efficiency gap: Implicit CoT provides the reasoning benefits of Explicit CoT without the latency and token costs.

Our findings validate that summarization planning can be successfully internalized through curriculum learning, establishing stepwise internalization as a viable training strategy beyond arithmetic and reasoning tasks. The substantial improvement over No-CoT (61% relative gain) confirms that implicit reasoning genuinely enhances summarization quality rather than adopting shortcuts. With inference efficiency 4× better than Explicit CoT, Implicit CoT enables reasoning-enhanced summarization in production systems where cost and latency matter.

Future work should explore larger models, additional domains, and hybrid approaches that selectively retain critical reasoning steps. As language models scale and deployment costs grow, techniques like Implicit CoT that preserve quality while reducing computational overhead will become increasingly important for practical NLP applications.

References

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Torsten Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17622–17630.

Zheng Chu and 1 others. 2024. Navigate through enigmatic labyrinth: A survey of chain-of-thought reasoning, advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Yuntian Deng, Yejin Choi, and Stuart Shieber. 2025. From explicit CoT to implicit CoT: Learning to internalize CoT step by step. In *Proceedings of the International Conference on Learning Representations*.

Hanjie Han, Yutong Xie, Hongyu Liu, Xiaohui Tang, Sayan Nag, William Headden, Cheng Luo, Yimeng Li, Shaoxiong Ji, Qian He, and Jiliang Tang. 2025. Reasoning with graphs: Structuring implicit knowledge to enhance LLMs reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of LLM agents: A survey. *arXiv preprint arXiv:2402.02716*.

Tianyi Lin, Juntao Xie, Shuaiyi Yuan, and Dong Yang. 2025. Implicit reasoning in transformers is reasoning through shortcuts. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Xuanhe Zhou, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, and 24 others. 2024. Tool learning with foundation models. *ACM Computing Surveys*, 57(4).

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS '23. Curran Associates Inc.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhi-Yuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. In *Advances in Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22. Curran Associates Inc.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, and 9 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Yang Xia and 1 others. 2025. A survey of chain-of-X paradigms for LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics.

Yiming Xu, Xin Guo, Zheng Zeng, and Chunyan Miao. 2025. Soft chain-of-thought for efficient reasoning with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3602–3622. Association for Computational Linguistics.