TORSO: Template-Oriented Reasoning Towards General Tasks

Minhyuk Kim, Seungyoon Lee, Heuiseok Lim[†]
Korea University, Republic of Korea
{mhkim0929, dltmddbs100, limhseok}@korea.ac.kr

Abstract

The approaches that guide Large Language Models (LLMs) to emulate human reasoning during response generation have emerged as an effective method for enabling them to solve complex problems in a step-by-step manner, thereby achieving superior performance. However, most existing approaches using few-shot prompts to generate responses heavily depend on the provided examples, limiting the utilization of the model's inherent reasoning capabilities. Moreover, constructing task-specific few-shot prompts is often costly and may lead to inconsistencies across different tasks. In this work, we introduce **T**emplate-**O**riented **R**ea**so**ning (TORSO), which elicits the model to utilize internal reasoning abilities to generate proper responses across various tasks without the need for manually crafted few-shot examples. Our experimental results demonstrate that TORSO achieves strong performance on diverse LLMs benchmarks with reasonable rationales.

1 Introduction

As the incorporation of human-like reasoning into Large Language Models (LLMs) has led to significant performance gains, numerous studies guiding the models to induce reasoning path via in-context learning approaches have been conducted (Chen et al., 2023; Wang et al., 2023; Wei et al., 2022). Most approaches include in-context learning methods such as Chain of Thought (CoT) (Wei et al., 2022) and Least-to-Most (LtM) (Zhou et al., 2023), which guide the model to follow specific reasoning paths through carefully curated few-shot prompts. On the other hand, training methods that leverage curated reasoning datasets based on given queries or instructions (OpenAI, 2024; Qwen, 2025; Guo et al., 2025) have become one of the specific approaches to elicit reasoning ability. These methods are fundamentally designed to expand the model's reasoning process, analogous to how humans approach complex problems step by step (Sun et al., 2024).

Such methods have shown their effectiveness in domains that require high-level reasoning, such as mathematics and science (Hwang et al., 2024). However, since in-context learning relies on imitating the reasoning process presented within the prompt, the design of the prompt plays a critical role in determining the model's performance (Stechly et al., 2024). Due to this dependency, designing prompts that yield consistent performance across various tasks remains a challenging problem (Zhang et al., 2025). Furthermore, building reasoning models trained with preconstructed rationales requires large amounts of additional training data, and they are often targeted at STEM domains, limiting their generalizability to a broader range of tasks (Bae et al., 2025).

In this work, we propose **T**emplate-**O**riented **R**ea**so**ning (TORSO), a method designed to guide LLMs to generate responses based on their inherent reasoning ability across various tasks, without relying on additional training data or task-specific few-shot prompts. TORSO is founded on the assumption that most LLMs already possess reasoning ability acquired through the vast amounts of training data they process during their learning phase. TORSO employs logit processing to guide the model's decoding process. We inject a specific token at the initiation stage to unlock reasoning ability and another at the end stage to encourage the model to wrap up the generated rationale into a final answer that directly aligns with the instruction.

Through the experiments on various LLM benchmarks, TORSO consistently outperforms the baselines across the tasks compared to in-context learning methods. Furthermore, we conduct a qualitative evaluation to assess the plausibility of the generated

[†]Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 16822–16830
November 4-9, 2025

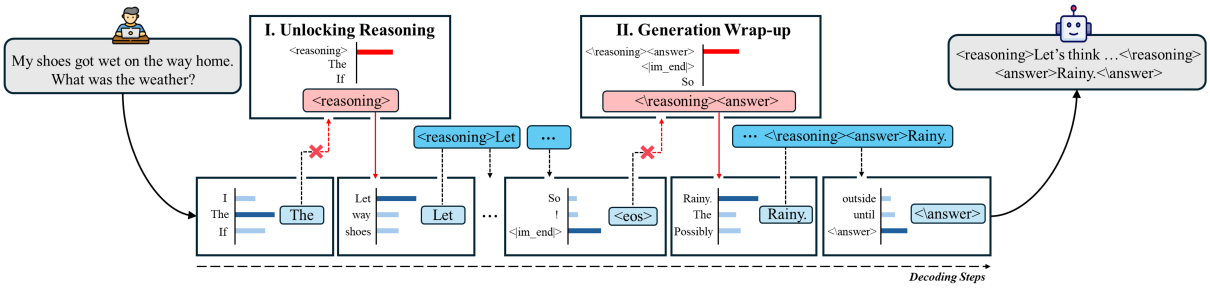

Figure 1: Overview of TORSO.

rationales. Our results indicate that the rationales induced by TORSO are appropriate. Our findings suggest that TORSO can effectively guide the models to construct their own reasoning paths without relying on few-shot prompts, and even demonstrate that conventional in-context learning approaches may hinder the coherent reasoning processes.

2 TORSO

TORSO is based on the assumption that LLMs inherently possess reasoning ability acquired during large-scale pretraining on sufficiently diverse and extensive data, even without explicit reasoning training. Based on this, we focus on eliciting the model's reasoning ability with minimal intervention. To this end, we manipulate the decoding process by forcibly adjusting the probability distribution over token generation. An overview of our pipeline is shown in Figure 1.

Step 1: Unlocking Reasoning We aim to activate the reasoning ability of the model by manipulating token generation probabilities. To be specific, we force the model to generate the token `<reasoning>` at the first step of decoding in response to a user query. Since LLMs generate outputs in an autoregressive manner, the presence of `<reasoning>` at the beginning of the sequence influences the subsequent generation process, encouraging the model to produce reasoning-oriented responses.

As shown in the Unlocking Reasoning part of Figure 1, we apply logit processing at the first decoding step of the LLM to forcibly assign very high logits to the tokens composing `<reasoning>`, regardless of the model's original probability distribution. This intervention ensures that the output sequence always begins with `<reasoning>`, providing the model with an explicit reasoning signal.

Step 2: Generation Wrap-up To ensure that the model produces a final answer that directly corresponds to the user's instruction following a long-form rationale, we inject a token that guides the model to refer its reasoning into a final answer. When the model is about to terminate generation by internally signaling an End-of-Sequence, we insert the `</reasoning>` token to indicate the end of the rationale, followed by the `<answer>` token to initiate answer generation.

This encourages the model to interpret the preceding output as a rationale. Consequently, the model incorporates the previously generated content when producing a final, synthesized answer. After generating the `<answer>` token, the model provides its final response and `</answer>` to close the answer field. Finally, internal signals end the decoding process by generating End-of-Sequence token.

By inducing the model to make a reasoning process itself, without constructing curated shots or training data for specific tasks, TORSO effectively harnesses the model's internal ability to form reasonable reasoning.

3 Experimental Setup

3.1 Implementation

Models For our experiments, we employ widely used open-source LLMs including Llama-3.1-8B-Instruct (Meta, 2024), Gemma-2-9B-it (Team et al., 2024), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023).

Hyperparameters To ensure consistent experimental conditions for all models, we apply the same decoding hyperparameters across all experiments. We set the maximum generation length to 8192 tokens, the temperature τ to 1.0, top-k to 50,

Model	Method	GSM8K	ARC	TruthfulQA	RACE	MMLU	GAOKAO	Avg.
DeepSeek-R1-Distill-Llama-8B	Base	0.8597	0.9226	0.4839	0.7744	0.6935	0.4982	0.7054
Llama-3.1-8B-Instruct	Base	0.7726	0.8493	0.4693	0.7145	0.5788	0.4090	0.6323
	CoT-Zero	0.7582	0.9125	0.4649	0.7625	0.5987	0.4496	0.6577
	CoT	0.7726	0.9162	0.5760	0.7493	0.6946	0.4369	0.6909
	ToT	0.7953	0.9082	0.4825	0.7570	0.6891	0.4126	0.6741
	LtM	0.7991	0.8923	0.5102	0.7507	0.6874	0.4551	0.6825
	TORSO	**0.8271**	**0.9301**	**0.5994**	**0.8440**	**0.7020**	**0.4759**	**0.7298**
gemma-2-9b-it	Base	0.7665	0.9263	0.6667	0.8572	0.6357	0.3714	0.7040
	CoT-Zero	0.7688	0.9668	0.7061	0.8684	0.7251	0.4842	0.7532
	CoT	0.7597	0.9482	0.7149	0.7911	0.7412	0.5237	0.7465
	ToT	0.7680	0.9680	0.7339	0.8538	0.7100	0.5200	0.7590
	LtM	0.7695	0.9689	0.7397	0.8809	0.7169	0.5455	0.7702
	TORSO	**0.8188**	**0.9705**	**0.7427**	**0.8893**	**0.7473**	**0.5474**	**0.7860**
Mistral-7B-Instruct-v0.2	Base	0.3942	0.7437	0.4254	0.6365	0.4106	0.2652	0.4793
	CoT-Zero	0.4147	0.7786	0.5351	0.6748	0.4993	0.2828	0.5309
	CoT	0.4193	0.7752	0.5190	0.6957	0.5172	0.3004	0.5378
	ToT	0.4215	0.7790	0.4942	0.6532	0.5127	0.3362	0.5328
	LtM	0.4071	0.7828	0.4401	0.6407	0.4793	0.3186	0.5114
	TORSO	**0.4375**	**0.7925**	**0.5906**	**0.7437**	**0.5541**	**0.3459**	**0.5774**

Table 1: Performance of different methods across various benchmarks, including average score.

and top-p to 1.0.

3.2 Evaluation

Quantitative To comprehensively evaluate the effectiveness of TORSO, we adopt six benchmark datasets targeting LLMs evaluation. These include GSM8K (Cobbe et al., 2021) and ARC-Easy (Clark et al., 2018) for mathematics and science, TruthfulQA (Lin et al., 2022) for evaluating reliability, RACE (Lai et al., 2017) for reading comprehension, and MMLU (Hendrycks et al., 2021) and GAOKAO (Zhong et al., 2023) for comprehensive knowledge evaluation. We assess model performance based on generated outputs using the Exact Match score as our primary evaluation metric.

Rationale Quality Evaluation We also conduct a pairwise qualitative evaluation to assess whether the induced reasoning from TORSO is appropriate to the given query. For the judge model, we employ GPT-4o. Samples that both TORSO and the corresponding baselines produce the correct answer are selected from the GSM8K dataset. For each case, the judge evaluates the rationale generated by TORSO and the baseline, determining which output provides a more suitable rationale for the given query.

To ensure a fair comparison with minimal positional bias, we randomly sample 200 questions and conduct 400 comparisons by reversing the order in which rationales are presented in the prompt. Each comparison is repeated five times per sample, and we report the win, tie, and lose ratios accordingly. The detailed prompt used for evaluation can be found in Appendix B.1.

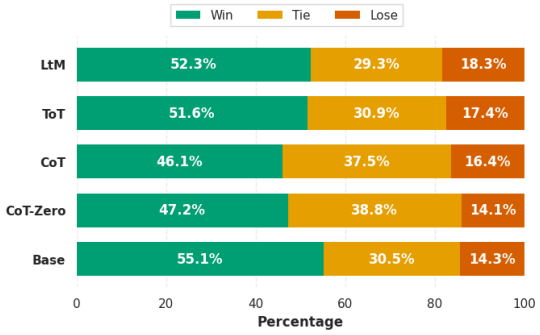

Figure 2: Qualitative comparison results of TORSO against baseline methods on the GSM8K benchmark using Llama-3.1-8B-Instruct. The bars represent the win, tie, and lose percentages.

3.3 Baselines

We adopt several major baselines, including the base model performance and representative in-context learning methods such as Chain-of-Thought (CoT)(Wei et al., 2022), Tree-of-Thought (ToT)(Yao et al., 2023), and Least-to-Most (LtM) (Zhou et al., 2023), all evaluated under the 5-shot setting. For CoT, we also consider a zero-shot CoT, where the phrase *"Let's think step by step."* is appended without providing any shots. In our experiments, ToT is applied in a few-shot in-context learning manner (Hulbert, 2023).

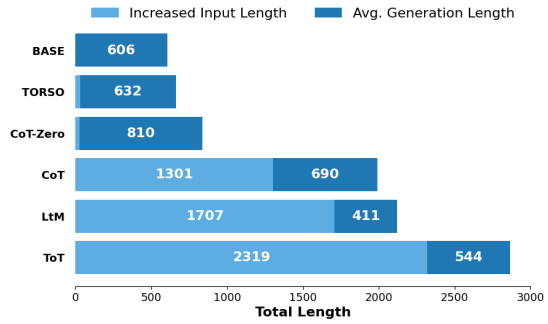

Figure 3: Increased input length and average generation length across methods. Each bar represents the increased input length (light blue) and average generation length (dark blue). Although TORSO does not increase the actual input length, we accounted for the forcibly appended tokens during decoding in our measurement.

4 Results

Overall Performance As shown in Table 1, TORSO consistently exceeds the baselines across the evaluated benchmarks. Specifically, TORSO remains effective even on MMLU and GAOKAO, which include a large number of questions focused on simple factual knowledge such as history and religion. This suggests that the existence of a reasoning can be significant even for questions that are relatively easy in difficulty or have a simple process for generating the correct answer. TORSO achieves these benefits through a minimal intervention, without the need for few-shot prompts or additional data construction.

When comparing CoT-zero and CoT, there are often cases where performance actually decreases when more shots are increased. Notably, for the gemma-2-9b-it model, CoT performs worse than CoT-zero on both GSM8K and ARC in mathematics and science where reasoning is generally known to be effective. This suggests that providing few-shot prompts to help the model reason may actually constrain the model's reasoning path to the shots.

We also observe that a discrepancy in performance can arise between models, depending on the methodology used for shot composition. For instance, with Mistral-7B-Instruct-v0.2, ToT yields better performance than LtM on all benchmarks except ARC, whereas gemma-2-9b-it performs better with LtM. However, TORSO consistently improves performance without relying on any shot composition, as it elicits the model's reasoning abilities through minimal intervention.

Comparison with Reasoning Model We also compare TORSO with existing reasoning model, DeepSeek-R1-Distill-Llama-8B. On all benchmarks except GSM8K and GAOKAO, TORSO applied to Llama 3.1 8B Instruct achieves superior performance. TORSO demonstrates higher performance even without learning additional reasoning data. This suggests that reasoning models, precisely because they focus on solving problems demanding STEM and complex logical structures, might be vulnerable on general benchmarks. This vulnerability highlights that reasoning models are not universally optimal across general tasks and underscores the value of TORSO.

Rationale Quality Analysis TORSO excels not only in its accuracy but also in the quality of the rationales it generates. Given that Figure 2 presents evaluations restricted to samples with correct answers, TORSO consistently achieves higher win rates across all cases, indicating that it resolves questions with more appropriate rationales. This suggests TORSO helps the model arrive at correct answers and produce reasonable rationales for solving the given tasks.

Length Efficiency TORSO is cost-efficient when considering both increased input and generation length. Figure 3 presents a comparison of the rationale lengths generated by each method when producing correct answers. All baselines that rely on few-shot prompting exhibit significant overhead in terms of input length.

In terms of generation length, CoT-zero produces the longest rationales, averaging 810.6 in length. Although CoT results in slightly shorter generations, this reduction is primarily due to the inclusion of few-shot prompts. In contrast, TORSO reaches correct answers with rationales that are comparably short to those generated by the base model. LtM and ToT yield generations with lengths of 411 and 544, respectively, which are shorter than TORSO's, but both methods require input lengths exceeding 1500. These results indicate that using a template to guide reasoning offers both high answer accuracy and improved efficiency, especially when compared to methods that rely on few-shot prompts.

5 Ablation Study

To better understand how the template used in TORSO affects model performance, we conduct

Category	Template	GSM8K	ARC	TQA	RACE
Base	–	0.773	0.849	0.469	0.715
CoT (5 shot)	–	0.773	0.916	0.576	0.695
TORSO (Ours)	<reasoning>+<answer>	**0.827**	0.930	**0.599**	**0.844**
Semantically Similar	<think>+<answer>	0.820	**0.936**	0.592	0.821
	<solution>+<answer>	0.814	0.911	0.585	0.826
	<reasoning>+<result>	0.826	0.927	0.599	0.831
	<reasoning>+<conclusion>	0.796	0.932	0.598	0.827
Arbitrary Placeholders	<partI>+<partII>	0.758	0.909	0.510	0.829
	<marker①>+<marker②>	0.749	0.840	0.518	0.742
Random Tokens	<xyz>+<abc>	0.724	0.848	0.500	0.662
	<qwer>+<asdf>	0.715	0.827	0.512	0.630

Table 2: Performance comparison of different forced token strategies across four benchmarks. TQA refers to TruthfulQA.

an ablation study using various templates. We evaluate performance under the same hyperparameter settings described in Section 3.1, using the Llama-3.1-8B-Instruct model across four benchmarks: GSM8K, ARC, TruthfulQA, and RACE. The results are presented in Table 2.

Semantically Similar In most configurations that replace the template used in TORSO with semantically similar templates (e.g., <think>, <solution>, <result>), performance surpasses the CoT (5-shot) baseline. These outcomes are consistent with the results observed when using the original <reasoning> + <answer> template in TORSO.

Arbitrary Placeholders In the case of arbitrary placeholders, the CoT (5-shot) baseline generally exceeds the performance. While the <part I > + <part II > template achieves a higher score on the RACE benchmark than the CoT (5-shot) baseline, this improvement does not consistently appear across other benchmarks. Although some specific tasks can benefit from arbitrary placeholders, their effectiveness is limited if their semantics do not directly contribute to the model's reasoning process.

Random Tokens Our results using templates of random tokens show lower performance than the base model across all benchmarks except TruthfulQA. This suggests that injecting meaningless templates during the decoding may interfere with the generation process.

These results indicate that the <reasoning> + <answer> template employed in TORSO consistently improves model performance. Additionally, replacing this template with semantically related alternatives can also lead to performance gains.

6 Conclusion

In this paper, we propose TORSO, a method that guides reasoning by enforcing a template during the model's decoding phase, without requiring additional effort such as constructing few-shot prompts or conducting training. Experiments across a diverse set of benchmarks demonstrate that TORSO effectively leverages the model's inherent reasoning ability to solve tasks and exhibits robust applicability across domains.

Limitations

While TORSO is an effective method for eliciting the reasoning abilities inherent in language models, it has several limitations. First, its applicability becomes challenging in scenarios that require reasoning over newly emerging information or tasks of extreme difficulty that fall outside the model's training distribution. This is because TORSO is primarily suited for inducing reasoning grounded in previously learned knowledge and patterns. Furthermore, reasoning models explicitly trained to generate rationales according to fixed templates may not reliably achieve performance gains when exposed to TORSO. In summary, TORSO is effective but limited in out-of-distribution reasoning tasks, highlighting the need for further investigation.

Ethics Statement

All experiments in this study were conducted with transparency and fairness, using only publicly available datasets intended for academic research. No personally identifiable information (PII) was involved. Furthermore, all models employed in this work are publicly released and accessible.

Acknowledgements

This work was supported by the Commercialization Promotion Agency for R&D Outcomes (COMPA) grant funded by the Korean government (Ministry of Science and ICT) (2710086166). This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korean government (MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2021R1A6A1A03045425).

References

Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, et al. 2025. Exaone deep: Reasoning enhanced language models. *arXiv e-prints*, pages arXiv–2503.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Preprint*, arXiv:2211.12588.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. *Preprint*, arXiv:2310.02226.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dave Hulbert. 2023. Using tree-of-thought prompting to boost chatgpt's reasoning. `https://github.com/dave1010/tree-of-thought-prompting`.

Hyeonbin Hwang, Doyoung Kim, Seungone Kim, Seonghyeon Ye, and Minjoon Seo. 2024. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards. *arXiv preprint arXiv:2404.10346*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *Preprint*, arXiv:2205.11822.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/meta-llama-3-1/`. Accessed: 2025-05-08.

OpenAI. 2024. Learning to reason with language models. `https://openai.com/index/learning-to-reason-with-llms/`. OpenAI Blog.

Qwen. 2025. Qwq-32b preview: Elevating mathematical and logical reasoning abilities of llms. `https://qwenlm.github.io/blog/qwq-32b-preview/`. Qwen Blog.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. Chain of thoughtlessness? an analysis of cot in planning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2024. A survey of reasoning with foundation models. *Preprint*, arXiv:2312.11562.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. *Preprint*, arXiv:2305.11792.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *arXiv preprint arXiv:2402.10200*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. 2023. Symbol tuning improves in-context learning in language models. *Preprint*, arXiv:2305.08298.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Xiang Zhang, Juntai Cao, Jiaqi Wei, Chenyu You, and Dujian Ding. 2025. Why does your cot prompt (not) work? theoretical analysis of prompt space complexity, its interaction with answer space during cot reasoning with llms: A recurrent perspective. *Preprint*, arXiv:2503.10084.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *Preprint*, arXiv:2411.14405.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

A Related Works

Prior work has explored in-context learning methods that provide human-like reasoning traces as part of the prompt, enabling models to imitate these reasoning processes when generating responses, particularly for solving relatively difficult tasks (Wei et al., 2022; Yao et al., 2023; Jung et al., 2022). Wang and Zhou (2024) attempt to intervene in the model's decoding path continuously, aiming to induce CoT reasoning without explicit prompting. However, such approaches do not fully leverage the model's inherent reasoning ability. Instead, they constrain the model to follow the reasoning path embedded in the prompt and ultimately reducing its behavior to mimicking the provided few-shot prompts (Wei et al., 2023).

On the other hand, there have been efforts to develop specialized reasoning models for solving high-difficulty reasoning tasks by directly injecting rationales into the model through large-scale training data (Zhao et al., 2024; Qwen, 2025; Guo et al., 2025; Bae et al., 2025). However, reasoning models that aim to maximize reasoning ability through supervised learning are typically targeted toward challenging STEM-related domains, which limits their problem-solving ability to those specific areas.

Inspired by Goyal et al. (2024), which inserts a token <pause> to induce a delay during both the training and inference stages of LLMs in order to enhance performance through inference time scaling, we propose TORSO, drawing inspiration from the idea that injecting specific tokens at inference can influence the entire decoding process.

B Qualitative Evaluation Details

B.1 Prompts

> Choose the better rationale for the given query. Answer with ①, ② or ③ for tie. Print only the answer.
> Query: [query]
> ① [a]
> ② [b]
> Answer:

We use special symbols ①,②, and ③ in the outputs generated by GPT-4o to facilitate the extraction of final judgements.

B.2 Total Results

Model	Comparison	Win	Tie	Lose
Llama-3.1-8B-Instruct	vs. BASE	1,103	610	287
	vs. CoT-Zero	944	775	281
	vs. CoT	922	751	327
	vs. ToT	1,033	619	348
	vs. LtM	1,047	587	366
gemma-2-9b-it	vs. BASE	829	663	508
	vs. CoT-Zero	812	616	572
	vs. CoT	714	602	684
	vs. ToT	897	521	582
	vs. LtM	824	498	678
Mistral-7B-Instruct-v0.2	vs. BASE	969	595	436
	vs. CoT-Zero	893	553	554
	vs. CoT	761	566	673
	vs. ToT	913	511	576
	vs. LtM	887	478	635

Table 3: Win/Tie/Lose counts from 2,000 pairwise qualitative comparisons (GPT-4o judge) of TORSO against each baseline on GSM8K rationales.

C Performance across MMLU Subject Categories

Model	Method	STEM	Other	Social Sciences	Humanities	Avg.
DeepSeek-R1-Distill-Llama-8B	Base	0.7697	0.7074	0.7319	0.6081	0.6935
Llama-3.1-8B-Instruct	Base	0.5503	0.6585	0.6324	0.5103	0.5788
	CoT-Zero	0.5779	0.6176	0.6737	0.5511	0.5987
	CoT	0.6689	0.7557	0.7673	0.6238	0.6946
	ToT	0.6343	0.7547	0.7943	0.6138	0.6891
	LtM	0.6318	0.7499	**0.7946**	0.6132	0.6874
	TORSO	**0.6803**	**0.7593**	0.7719	**0.6329**	**0.7020**
gemma-2-9b-it	Base	0.6629	0.6682	0.6991	0.5547	0.6357
	CoT-Zero	0.7187	0.7612	0.8122	0.6487	0.7251
	CoT	**0.7371**	0.7741	0.8222	0.6693	0.7412
	ToT	0.6977	0.7361	0.7975	0.6438	0.7100
	LtM	0.7035	0.7602	0.8047	0.6400	0.7169
	TORSO	0.7279	**0.7773**	**0.8278**	**0.6878**	**0.7473**
Mistral-7B-Instruct-v0.2	Base	0.3781	0.4026	0.4524	0.4102	0.4106
	CoT-Zero	0.4738	0.5356	0.5388	0.4665	0.4993
	CoT	0.4554	0.5645	0.5967	0.4752	0.5172
	ToT	0.4932	0.5220	0.5596	0.4891	0.5127
	LtM	0.4164	0.5233	0.5525	0.4444	0.4793
	TORSO	**0.4941**	**0.6241**	**0.6282**	**0.4995**	**0.5541**

Table 4: Breakdown of MMLU performance by subject category.

D Qualitative Examples of Reasoning

Model	Llama-3.1-8B-Instruct
Task	ARC-Easy
Question	A research scientist writes a paper on the initial regrowth of a forest after a fire has damaged the entire ecosystem. Which title would be best for the paper? A: Primary Succession B: Stable Communities C: Climax Communities D: Secondary Succession
Answer (TORSO)	<reasoning> The correct title for the paper would be related to the process of regrowth after a fire has damaged the entire ecosystem. Secondary succession (D) is the process of regrowth and establishment of a new ecosystem in an area that was previously inhabited but has been disturbed or damaged, such as after a fire or clear-cutting. Given the context of the paper, the scientist is studying the regrowth of a forest after a fire, which indicates that the ecosystem was previously inhabited but was damaged. Therefore, the correct title would be related to secondary succession. </reasoning> <answer>D</answer>
Wrong Answer (CoT)	A forest fire has completely destroyed the ecosystem. That means all the plants and animals are gone, and the land is now barren. When life starts to grow again in an area with no organisms or soil, it is called primary succession. Since the forest is starting from scratch after the fire, this situation fits the definition of primary succession. Therefore, the best title for the paper would be: A.

Model	Mistral-7B-Instruct-v0.2
Task	RACE
Question	For most people, the word "fashion" means "clothes". But people may ask the question, "What clothes are in fashion?" And they use the word "fashionable" in the same way. "She was wearing a fashionable coat. His shirt was really a fashionable colour." But of course there are fashions in many things, not only in clothes. There are fashions in holidays, in restaurants, in films and books. There are even fashions in school subjects, jobs ... and in languages. Fashions change as time goes. If you look at pictures of people or things from the past, you will see that fashions have always changed. An English house of 1750 was different from one of 1650. A fashionable man in 1780 looked very different from his grandson in 1860. Today fashions change very quickly. Some of this is natural. We hear about things much more quickly than in the past. Newspapers, radios, telephones and television send information from one country to another in a few hours. New fashions mean that people will buy new things, so you see _ . By _ we can see that fashions change as time goes. A: finding a job B: looking at pictures C: eating in a restaurant D: wearing a fashionable coat
Answer (TORSO)	<reasoning> The context of the text suggests that the blank should be filled with an action that can be associated with observing or discovering fashions. Among the given options, "looking at pictures" is the most suitable answer as it aligns with the idea of examining past and present fashions. The text discusses how fashions change over time and can be observed in various aspects of life, including clothes, holidays, restaurants, films and books, school subjects, jobs, and even languages. The author mentions that people can look at pictures to see how fashions have evolved. Therefore, the correct answer is "looking at pictures." </reasoning> <answer>B</answer>
Wrong Answer (CoT)	The passage begins by explaining that many people associate fashion with clothing, and it gives specific examples like a fashionable coat and a fashionable shirt color. Since these examples are mentioned early in the text, it seems like they are important to understanding how fashion works. Wearing a fashionable coat is a direct expression of current fashion trends, so by observing people who wear such clothing, we can understand what is considered fashionable at a certain time. If someone is wearing a coat that matches the popular styles of the time, it reflects the fashion of that era. Therefore, by wearing a fashionable coat, we can see how fashions change as time goes. This makes it a reasonable answer to complete the sentence in the passage. So the answer is D.