D-ProtoCoT: Prototype-Based Path Selection for Chain-of-Thought Reasoning

Anonymous ACL submission

Abstract

Chain-of-Thought (CoT) prompting enables large language models to solve complex reasoning tasks by generating explicit multi-step reasoning paths. To improve robustness, self-consistency aggregates multiple sampled CoT paths via majority voting over final answers. However, this strategy operates purely at the answer level and implicitly assumes uniform reasoning quality, overlooking substantial variation among reasoning paths that yield the same answer.

In this work, we propose **D-ProtoCoT**, an inference-time framework for selecting high-quality reasoning paths based on representation-level alignment. D-ProtoCoT first constructs a contrastively aligned embedding space using weak supervision from gold answer matching, in which representations of correct reasoning paths are encouraged to cluster while incorrect ones are separated. At inference time, a dynamic reasoning prototype is formed by aggregating multiple sampled paths in this space, and the reasoning path most aligned with the prototype is selected.

D-ProtoCoT does not modify the underlying language model and requires no additional annotation beyond gold answers for representation alignment. Experiments on CommonsenseQA, GSM8K, and StrategyQA with LLaMA-3.1-8B-Instruct and Qwen3-8B show that D-ProtoCoT consistently outperforms self-consistency across most settings, achieving up to 23.6% absolute improvement on StrategyQA. Further analysis demonstrates that representation alignment is essential, as naïve centroid-based selection with frozen embeddings yields substantially inferior performance, indicating that semantic alignment provides a more reliable signal than answer frequency for reasoning path selection.

1 Introduction

Large language models (LLMs) have shown strong performance on multi-step reasoning tasks when guided by Chain-of-Thought (CoT) (Wei et al., 2023) prompting, which encourages models to generate explicit intermediate reasoning steps before producing a final answer. To further improve robustness, self-consistency samples multiple CoT reasoning paths and aggregates their final answers via majority voting. This combination has become a standard inference strategy for reasoning benchmarks such as GSM8K, CommonsenseQA, and StrategyQA.

Despite its effectiveness, self-consistency operates solely at the answer level. It implicitly assumes that all reasoning paths leading to the same answer are equally reliable, and that answer frequency is a sufficient proxy for reasoning quality. However, in practice, different reasoning paths may arrive at the same answer through substantially different logical processes, including flawed or internally inconsistent reasoning. As a result, answer-level agreement often fails to reflect the true quality of reasoning.

This limitation naturally raises the question of whether reasoning paths can be evaluated and compared directly, rather than aggregated only through their final answers. A straightforward idea is to embed reasoning chains into a semantic space and perform aggregation using simple statistics such as centroids. However, empirical analysis shows that such naïve semantic aggregation is insufficient. Even with strong pretrained encoders, correct and incorrect reasoning paths are often entangled in the embedding space, as surface-level semantic similarity does not reliably capture logical validity.

This work proposes **D-ProtoCoT** (Dynamic Prototype-based Chain-of-Thought), an inference-time framework for reasoning path selection that operates in a contrastively aligned representation space. The core idea of D-ProtoCoT is to explicitly align reasoning representations such that correct reasoning paths are pulled closer to their corresponding questions, while incorrect paths are

pushed away. This alignment is achieved by fine-tuning a lightweight encoder with a contrastive objective derived from gold answer matching, without modifying the underlying language model.

Building on this aligned space, D-ProtoCoT constructs a *dynamic reasoning prototype* for each question by aggregating multiple sampled reasoning paths with similarity-based weighting. Rather than relying on majority voting, the framework selects the reasoning path that is most aligned with the prototype, thereby prioritizing internally coherent reasoning trajectories. All selection is performed at inference time, and no additional supervision or model updates are required beyond encoder alignment.

D-ProtoCoT is evaluated on GSM8K, CommonsenseQA, and StrategyQA using two representative instruction-tuned LLMs: LLaMA-3.1-8B-Instruct and Qwen3-8B. Experimental results show that D-ProtoCoT consistently outperforms standard CoT decoding and self-consistency across tasks and model families, with particularly large gains on reasoning-intensive benchmarks. Further ablation studies and qualitative analyses demonstrate that representation alignment is critical, as naïve centroid-based selection with frozen embeddings yields substantially inferior performance.

The main contributions of this work are summarized as follows:

- D-ProtoCoT is proposed as an inference-time framework for selecting chain-of-thought reasoning paths based on representation-level alignment rather than answer-level aggregation.
- A contrastive representation alignment mechanism is introduced to enable the construction of dynamic reasoning prototypes, without modifying or fine-tuning the underlying language model.
- The effectiveness of D-ProtoCoT is demonstrated across multiple reasoning benchmarks and model families, with analysis highlighting its advantage over naïve semantic aggregation methods.

2 Related Work

2.1 Chain-of-Thought Reasoning and Self-Consistency

Chain-of-Thought (CoT) prompting enables large language models to solve multi-step reasoning problems by explicitly generating intermediate reasoning steps (Wei et al., 2023). Building on this paradigm, self-consistency improves robustness by sampling multiple reasoning paths and aggregating final answers through majority voting (Wang et al., 2023). This strategy has become a standard inference-time technique for reasoning benchmarks such as GSM8K, CommonsenseQA, and StrategyQA.

Despite its empirical success, self-consistency operates purely at the answer level and implicitly assumes that reasoning paths leading to the same answer are equally reliable. Several studies have noted that majority voting may amplify spurious reasoning patterns and fail when correct reasoning paths are not numerically dominant (Fu et al., 2023; Zhou et al., 2022). D-ProtoCoT departs from answer-level aggregation by directly operating on reasoning representations, enabling path selection based on semantic and structural alignment rather than answer frequency.

2.2 Reasoning Path Selection and Verification

A related line of work focuses on selecting or validating reasoning paths using external signals. Verifier-based approaches train separate models to assess the correctness of generated reasoning chains or intermediate steps (Cobbe et al., 2021; Lightman et al., 2023). Similarly, reward-model-based methods employ reinforcement learning or preference optimization to encourage correct reasoning trajectories (Ouyang et al., 2022; Bai et al., 2022).

While effective, these approaches typically require additional supervision, costly human annotation, or task-specific training, and often modify the base language model. In contrast, D-ProtoCoT is designed as a lightweight inference-time framework. It does not attempt to explicitly judge logical correctness, nor does it update the reasoning model. Instead, it relies on representation-level alignment to identify reasoning paths that are semantically consistent with high-quality reasoning patterns, offering a complementary alternative to explicit verification.

2.3 Contrastive Learning for Reasoning Representations

Contrastive learning has been widely adopted to improve semantic representation quality in natural language processing, particularly for sentence and text embedding models (Gao et al., 2021; Gunel

et al., 2021). Recent work has explored contrastive objectives for reasoning-related tasks, including contrastive chain-of-thought prompting (Chia et al., 2023) and reasoning-aware representation learning for logical reasoning problems (Yu et al., 2020).

Unlike prior approaches that apply contrastive learning directly to language model outputs or confidence estimation, D-ProtoCoT employs contrastive alignment to explicitly shape a shared embedding space for questions and reasoning paths. Weak supervision from gold answer matching is used to distinguish correct and incorrect reasoning paths, enabling the construction of an aligned representation space without additional annotation or model modification. This aligned space serves as the foundation for prototype-based reasoning path selection.

2.4 Prototype-Based Representation Modeling

Prototype-based methods have a long history in representation learning and pattern recognition, where class prototypes or centroids are used to capture central tendencies of data distributions (Snell et al., 2017). In NLP, prototype-based representations have been applied to text classification, clustering, and semantic matching tasks (Han et al., 2018; Pan et al., 2020).

Recent studies suggest that prototype aggregation can improve robustness by mitigating noise from individual samples (Liu et al., 2022). D-ProtoCoT extends this idea to reasoning path selection by constructing a dynamic, question-specific prototype at inference time. Unlike static class prototypes, the proposed dynamic prototype aggregates multiple reasoning paths with similarity-based weighting, allowing the model to adaptively capture the dominant semantic structure of high-quality reasoning for each input instance.

3 Method

3.1 Problem Setup

Given an input question q, a large language model (LLM) generates multiple chain-of-thought (CoT) reasoning paths $\mathcal{C} = \{c_1, \ldots, c_K\}$ via stochastic decoding. Each reasoning path consists of a sequence of intermediate reasoning steps followed by a final answer. The objective is to select a single reasoning path that exhibits strong internal coherence and leads to a reliable final answer.

Throughout this framework, the base language model used for reasoning path generation remains frozen. Only a lightweight auxiliary encoder is trained for representation alignment. All components operate at inference time, enabling reasoning path selection without modifying or fine-tuning the underlying language model.

3.2 Multi-Granular Representation Extraction

Each reasoning path is encoded using a BERT-base-uncased encoder $E(\cdot)$, which is subsequently fine-tuned through contrastive representation alignment (Section 3.3). Given the tokenized input of a reasoning path, the encoder produces a sequence of hidden states from its final transformer layer.

To capture reasoning information at different granularities, representations are extracted at three levels.

Token-Level Representation. Let $\mathbf{H}_i \in \mathbb{R}^{T \times d}$ denote the final-layer hidden states for path c_i, where T is the number of tokens and d is the hidden dimension. These token embeddings provide the most fine-grained representation of the reasoning process.

Step-Level Representation. Each reasoning path c_i is segmented into M steps $\{u_{i,1}, \ldots, u_{i,M}\}$ based on newline delimiters. Let $\mathcal{T}_{i,j} \subseteq \{1, \ldots, T\}$ denote the set of token indices belonging to step $u_{i,j}$. The step-level representation is computed as:

$$\mathbf{s}_{i,j} = \frac{1}{|\mathcal{T}_{i,j}|} \sum_{t \in \mathcal{T}_{i,j}} \mathbf{h}_{i,t},$$

where $\mathbf{h}_{i,t} \in \mathbb{R}^d$ is the token embedding at position t.

Path-Level Representation. A single representation for the entire reasoning path is obtained by mean pooling over its step embeddings:

$$\mathbf{z}_i = \frac{1}{M} \sum_{j=1}^{M} \mathbf{s}_{i,j}.$$

The resulting path-level embedding $\mathbf{z}_i \in \mathbb{R}^d$ serves as the primary unit for contrastive alignment and prototype-based selection.

3.3 Contrastive Representation Alignment

Motivation. naïve semantic representations obtained from off-the-shelf encoders often fail to distinguish logically valid reasoning from superficially plausible but incorrect chains, as they are

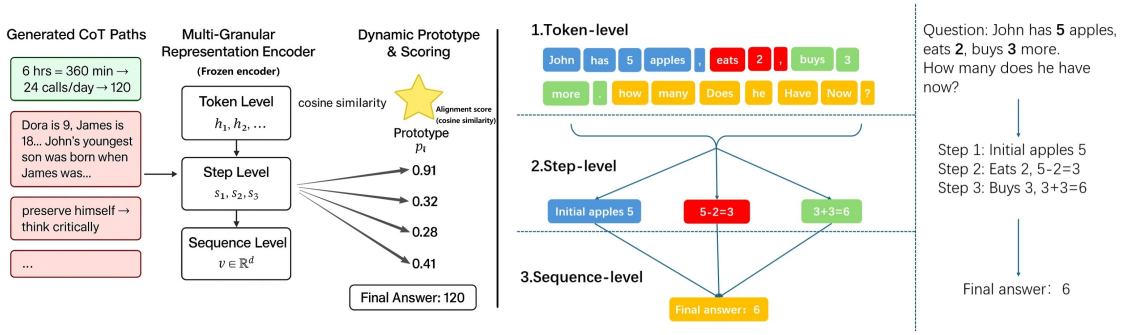

Figure 1: Overview of D-ProtoCoT. The framework encodes multiple chain-of-thought reasoning paths into a contrastively aligned representation space, constructs a dynamic reasoning prototype, and selects the path with the highest prototype alignment score at inference time.

not explicitly trained to align questions with valid reasoning processes. To address this limitation, a contrastive learning stage is introduced to shape the representation space such that correct reasoning paths are pulled closer to their corresponding questions, while incorrect paths are pushed away. This alignment ensures that downstream prototype construction operates in a space where semantic proximity correlates with reasoning quality.

Positive and Negative Path Construction. For each question q, K reasoning paths are generated and labeled as correct or incorrect based on exact match between the final answer and the ground-truth answer. Paths whose final answer matches the gold label are treated as positive samples, while all remaining paths are treated as negatives. This supervision signal is obtained automatically and requires no additional annotation.

Representation Encoding. The question q and each reasoning path c_i are encoded using the same encoder $E(\cdot)$. Let $\mathbf{z}_q = E(q)$ denote the pooled question representation, and $\mathbf{z}_i = E(c_i)$ denote the path-level representation. All encoder parameters are fine-tuned during contrastive alignment.

Contrastive Objective. An InfoNCE-style contrastive loss is employed to align question and reasoning representations:

$$\mathcal{L}_{\text{con}} = -\log \frac{\sum_{i \in \mathcal{P}} \exp\left(\text{sim}(\mathbf{z}_q, \mathbf{z}_i)/\tau\right)}{\sum_{j \in \mathcal{P} \cup \mathcal{N}} \exp\left(\text{sim}(\mathbf{z}_q, \mathbf{z}_j)/\tau\right)},$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the sets of correct and incorrect reasoning paths, $\text{sim}(\cdot, \cdot)$ is cosine similarity, and τ is a temperature hyperparameter.

3.4 Dynamic Prototype Construction

After contrastive alignment, reasoning paths are embedded into a representation space where correct paths are encouraged to align more closely with their corresponding questions. Based on this space, a dynamic reasoning prototype is constructed for each question at inference time.

Given a question q and its K reasoning paths with embeddings $\{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$, the dynamic prototype $\mathbf{p}_q$ is defined as a weighted combination of all path representations:

$$\mathbf{p}_q = \sum_{i=1}^{K} w_i \mathbf{z}_i.$$

The weight w_i reflects the alignment between a reasoning path and the question representation:

$$w_i = \frac{\exp(\text{sim}(\mathbf{z}_q, \mathbf{z}_i))}{\sum_{j=1}^{K} \exp(\text{sim}(\mathbf{z}_q, \mathbf{z}_j))}.$$

This similarity-based weighting assigns higher influence to paths that are more aligned with the question, while retaining information from all sampled paths.

3.5 Prototype-Based Alignment and Path Selection

Given the dynamic prototype $\mathbf{p}_q$, an alignment score is computed for each reasoning path using cosine similarity:

$$s_i = \text{sim}(\mathbf{z}_i, \mathbf{p}_q).$$

The final reasoning path is selected by maximizing the alignment score:

$$c^* = \arg\max_{c_i \in \mathcal{C}} s_i.$$

The final answer is extracted from the selected top-1 reasoning path c^*.

3.6 Inference-Time Properties

D-ProtoCoT operates entirely at inference time. No parameters are updated during reasoning path selection, and the framework introduces minimal computational overhead beyond reasoning generation and embedding computation. The method is compatible with any language model capable of generating chain-of-thought reasoning paths.

4 Experiments

This section evaluates the effectiveness of D-ProtoCoT on three representative multi-step reasoning benchmarks: GSM8K, CommonsenseQA, and StrategyQA. These datasets emphasize arithmetic reasoning, commonsense inference, and implicit semantic reasoning respectively, enabling a comprehensive assessment of reasoning path selection under diverse reasoning characteristics.

4.1 Benchmarks and Backbone Models

D-ProtoCoT is evaluated using two instruction-tuned large language models: LLaMA-3.1-8B-Instruct and Qwen3-8B. These models differ substantially in pretraining data, architectural design, and instruction-following behavior, enabling an assessment of robustness and generalization across heterogeneous model families.

4.2 Reasoning Path Generation

For each input question, **10** independent chain-of-thought reasoning paths are sampled using temperature-controlled decoding. The same set of sampled reasoning paths is shared across all compared methods to ensure fair comparison. Final answers are extracted from each reasoning path following the official evaluation protocols of each benchmark.

4.3 Baselines

D-ProtoCoT is compared with standard inference-time reasoning baselines:

Standard CoT selects the answer from a single sampled reasoning path.

Self-Consistency aggregates multiple reasoning paths via majority voting over final answers.

C-CoT (Chia et al., 2023) selects reasoning paths based on confidence estimation.

4.4 Implementation Details

Reasoning paths are encoded using a pretrained BERT encoder. Token-level representations are extracted from the **final transformer layer**. Step-level representations are obtained by mean-pooling token embeddings within each reasoning step, while path-level representations are computed by mean-pooling all step embeddings of a reasoning path, following the representation definitions in Section 3.2.

For D-ProtoCoT, the encoder is contrastively aligned using reasoning paths with weak supervision derived from gold answer matching (exact match), which identifies positive and negative reasoning paths. The contrastive objective fine-tunes the entire encoder, rather than a separate projection head. During inference, all reasoning paths are encoded using the aligned encoder, and a dynamic prototype is constructed via similarity-weighted aggregation of all paths, where weights are determined by their alignment with the question representation.

All experiments are conducted on a single Nvidia RTX 4090 GPU. All methods use identical generation settings, the same number of sampled reasoning paths, and the same evaluation protocol. No additional fine-tuning is performed on the backbone LLMs.

4.5 Main Results

Table 1 shows that D-ProtoCoT consistently outperforms standard CoT decoding and achieves competitive or superior results compared to self-consistency across most settings. Improvements are particularly pronounced on tasks where correct reasoning paths do not dominate numerically, indicating that representation-level reasoning alignment provides a complementary signal beyond answer frequency.

4.6 Inference Cost

D-ProtoCoT introduces modest inference overhead due to multi-path encoding and prototype scoring. However, the dominant cost remains in LLM generation. The additional representation encoding and similarity computation incur negligible latency in practice. Empirically, D-ProtoCoT exhibits comparable inference time to self-consistency when using the same number of sampled paths.

Table 1: Main results (%) on CommonsenseQA, GSM8K, and StrategyQA. Results are reported using 10 sampled reasoning paths.

	LLaMA-3.1-8B-Instruct			Qwen3-8B		
Method	CSQA	GSM8K	StrategyQA	CSQA	GSM8K	StrategyQA
Standard CoT	68.23	43.80	68.60	75.60	90.40	66.80
Self-Consistency	77.40	63.20	62.60	86.98	90.65	**76.60**
C-CoT	64.40	**87.00**	64.40	68.20	75.60	62.70
D-ProtoCoT	**79.80**	78.20	**86.20**	**87.71**	**94.15**	72.60

5 Analysis

D-ProtoCoT is designed as an inference-time reasoning path selection framework, rather than a training-time modification of language models. This design choice prioritizes modularity and ease of integration, allowing the method to be applied to a wide range of pretrained models without additional fine-tuning or supervision.

Unlike verifier-based or reward-model-based approaches, D-ProtoCoT does not explicitly judge logical correctness. Instead, it relies on representation-level alignment to identify reasoning paths that are structurally and semantically consistent with high-quality reasoning patterns. This makes D-ProtoCoT a lightweight middleware situated between naïve aggregation strategies and heavy verification pipelines.

5.1 Why naïve Centroid Selection Fails

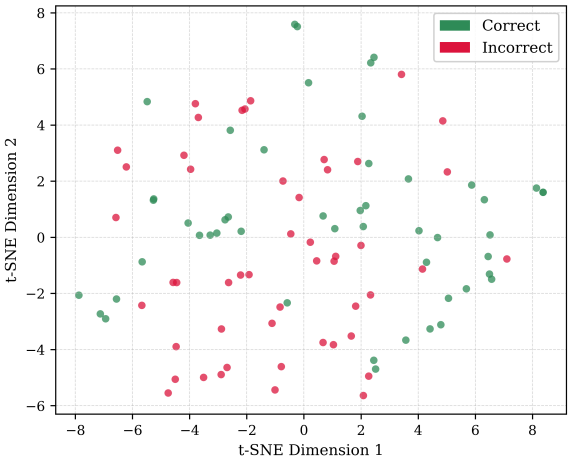

Figure 2: t-SNE visualization of BERT embeddings from 100 chain-of-thought paths generated for a StrategyQA example. Green points indicate correct reasoning paths, while red points indicate incorrect ones. Strong overlap suggests that surface-level semantic similarity alone is insufficient to distinguish reasoning quality.

A natural baseline for reasoning path selection is to compute embeddings for each reasoning path and select the one closest to the centroid of all paths. As shown in Figure 2, frozen BERT embeddings of correct and incorrect reasoning paths are heavily intermixed, indicating that surface-level semantic similarity is insufficient to separate logically valid reasoning from flawed ones. This limitation is further confirmed quantitatively in Table 2, where naïve centroid-based selection with raw BERT embeddings performs poorly compared to D-ProtoCoT.

Table 2: Comparison on GSM8K (LLaMA-3.1-8B-Instruct) between naïve centroid-based selection and D-ProtoCoT.

Method	**Accuracy (%)**
Raw BERT + Centroid	21.00
D-ProtoCoT (ours)	**78.20**

This result demonstrates that without representation alignment, centroid-based aggregation is vulnerable to selecting semantically similar but logically incorrect reasoning paths. In contrast, D-ProtoCoT benefits from contrastively aligned representations that better reflect reasoning coherence.

5.2 Case Study

Figure 3 illustrates a representative failure case of self-consistency. Although the majority of reasoning paths lead to the same incorrect answer, D-ProtoCoT selects a path with higher internal coherence by aligning with the reasoning prototype.

5.3 Effect of Representation Granularity

D-ProtoCoT encodes reasoning paths at multiple levels of granularity. Token-level representations capture surface variations, while path-level representations summarize global semantics but may obscure localized logical errors. Step-level representations provide an intermediate abstraction by

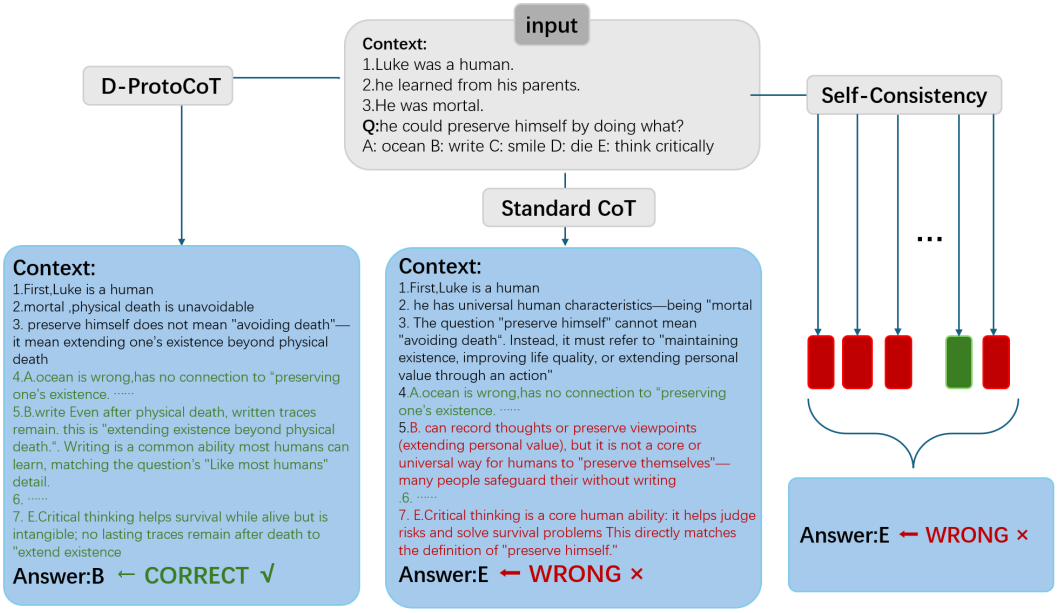

Figure 3: Case study on StrategyQA. While self-consistency selects an incorrect answer due to majority voting, D-ProtoCoT identifies a reasoning path that correctly interprets the implicit semantics of the question.

modeling semantically meaningful inference units.

Although we do not present separate quantitative experiments for each granularity, qualitative analysis suggests that step-level representations are particularly effective for multi-step reasoning tasks, where errors often arise from a single flawed intermediate step.

5.4 Impact of Backbone Model Strength

We observe that when the backbone model is a strong instruction-tuned LLM such as Qwen3-8B, self-consistency can already achieve competitive performance with a limited number of sampled reasoning paths. In this setting, the performance gap between self-consistency and D-ProtoCoT narrows.

This observation suggests that the benefit of prototype-based reasoning selection is most pronounced when reasoning path quality exhibits higher variance, or when a larger number of candidate reasoning paths are available.

5.5 Scope and Limitations

D-ProtoCoT is not intended to replace formal verification or symbolic reasoning systems. Instead, it serves as a lightweight and modular component that can be easily integrated into existing reasoning pipelines. Its effectiveness depends on the diversity and quality of sampled reasoning paths, and its benefits diminish when most paths are already coherent and correct.

6 Conclusion

This work investigates inference-time reasoning path selection for large language models beyond answer-level aggregation. To this end, D-ProtoCoT is proposed—a prototype-based framework that selects chain-of-thought paths by measuring their alignment with a dynamically constructed reasoning prototype in a contrastively aligned representation space. By prioritizing internal semantic coherence over surface-level answer agreement, the approach offers a complementary strategy to existing decoding and aggregation methods.

Empirical results across multiple reasoning benchmarks and model families demonstrate that D-ProtoCoT consistently improves performance in settings where majority voting is suboptimal, particularly when correct reasoning paths are semantically concentrated while incorrect ones are scattered. Further analysis highlights the insufficiency of naïve semantic aggregation and underscores the importance of structured representation learning for reliable reasoning path selection.

Although D-ProtoCoT does not guarantee logical correctness in all cases, it provides evidence that representation-level alignment serves as a more informative signal than answer frequency alone. Future work may explore finer-grained reasoning representations, integration with formal verification, or adaptation to settings without gold answers.

Limitations

D-ProtoCoT has several limitations that warrant discussion. First, the method assumes access to gold answers during the *training phase* of the auxiliary encoder for contrastive alignment (3.3). Consequently, it is not directly applicable to fully unsupervised settings where no labeled answers are available, which may limit its use in certain real-world deployment scenarios.

Second, the effectiveness of reasoning path selection depends on the diversity and correctness distribution of the sampled reasoning paths. When most generated paths are incorrect or semantically homogeneous (e.g., due to low decoding temperature), the dynamic prototype may fail to capture a meaningful signal, thereby reducing potential performance gains. This observation is consistent with our empirical findings that D-ProtoCoT provides smaller improvements on tasks with inherently low path diversity.

Third, D-ProtoCoT introduces additional inference overhead due to the use of an auxiliary encoder for representation extraction and similarity computation. Although this overhead is modest compared to the cost of large language model generation, it still results in higher latency than vanilla self-consistency. Moreover, the method may inherit biases or domain mismatches from the encoder's pretraining data.

Finally, all experiments are conducted on English-language benchmarks. The effectiveness of D-ProtoCoT on low-resource languages or specialized domains, such as medical or legal reasoning, remains unexplored and warrants further investigation.

Acknowledgments

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *Preprint*, arXiv:2311.09277.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. *Preprint*, arXiv:2210.00720.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022. Learn from relation information: Towards prototype representation rectification for few-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1822–1831, Seattle, United States. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. *Preprint*, arXiv:2004.12704.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *NeurIPS*, pages 4077–4087.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and

Denny Zhou. 2023. 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. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *Preprint*, arXiv:2002.04326.

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