OntoCoT: Efficient Domain Alignment of LLMs via Ontology-Guided Chain-of-Thought Fine-tuning

Anonymous ACL submission

Abstract

Domain-specific expert models typically rely on large-scale domain corpora for fine-tuning. As constructing such datasets is very costly, previous works have introduced ontologies as a data source. These approaches often focus on aligning models with domain knowledge, while overlooking the alignment of models with domain reasoning. Inspired by chain-of-thought distillation and the mind-map-like characteristics of ontologies, we propose **OntoCoT**, a framework that introduces Chain-of-Thought (CoT) as a knowledge carrier to guide model reasoning. By leveraging the hierarchical and graph-structured properties of ontologies, OntoCoT enhances the model's ability to organize and utilize domain knowledge, while also improving its efficiency in learning new knowledge. We further design a selection strategy to filter CoT sequences that are highly informative, high-quality, and underutilized by the model. Relying solely on existing ontologies and LLMs, our framework reduces the cost of data construction and maintenance, improves generalization ability, and achieves significant performance gains compared to previous methods even when using small ontology datasets, making it especially suitable for domains with limited ontology resources.

1 Introduction

In recent years, to enhance the capabilities of large models in specific domains such as finance and biology, both academia and industry have investigated domain-specific LLMs (Liu et al., 2023; Labrak et al., 2024). However, previous approaches often rely on large-scale domain-specific text corpora, involving high construction costs, fragmented knowledge, and difficulties in collecting data due to privacy concerns within certain domains.

To address this, previous works have introduced widely established and rigorously constructed ontologies as a data source. For example, TaxoLlama(Moskvoretskii et al., 2024) directly employs the WordNet ontology for model fine-tuning, while OntoTune (Liu et al., 2025) generates knowledge-card-like data as training corpora. However, these approaches often focus solely on aligning models with domain knowledge, by generating large amounts of static knowledge text, such as knowledge cards or domain introductions. They neglect aligning models with domain reasoning, i.e., the model's ability to effectively apply the knowledge for reasoning over domain-specific problems, which significantly limits its capacity to organize and utilize knowledge in practical scenarios.

Inspired by chain-of-thought distillation, which leverages intermediate reasoning traces to improve reasoning capabilities (Zelikman et al., 2022; Kojima et al., 2022), we observe that ontologies possess mind-map-like characteristics, motivating the introduction of chain-of-thought as a knowledge carrier. However, how to efficiently generate such chain-like data based on ontologies and how to effectively filter it to obtain high-quality sequences remain challenges that require further investigation.

Thus, we propose an ontology-guided thought chain framework for model self-evolution. Our goal is to use ontology knowledge to prompt the LLM to construct thought chains and integrate the ontology relationships centered around target entities into corresponding thought chains. By fine-tuning the LLM with these Chain-of-Thought techniques, we encourage the model to better utilize the existing knowledge and associate ontology knowledge when answering domain-specific questions.

Our workflow consists of three main steps: **(1) Associative CoT Instruction Generation**: We use open-ended thinking instructions to generate multiple thought chains, both with and without ontology knowledge prompts. **(2) High-quality Instruction Text Selection**: We combine various methods to analyze model mastery, knowledge richness, and thought chain quality, selecting high-quality thought chains for training. **(3)**

LLM Fine-tuning: We use the selected training set to fine-tune the base model, resulting in the final domain-specific LLM. We first conduct our study in the medical field, using the high-quality medical ontology SNOMED CT (Donnelly, 2006) as the source ontology. To evaluate the effectiveness of our method, we compare it to existing domain LLMs trained on large-scale corpora, as well as the ontology-based method TaxoLLaMA (Moskvoretskii et al., 2024) and OntoTune (Liu et al., 2025) implemented on the same LLM. The results show that our method achieves state-of-the-art performance in out-of-ontology task domain question-answering (QA), demonstrating that it can effectively enhance domain-specific task capabilities. Furthermore, compared to previous ontology-based works, our method uses much fewer ontology samples (i.e., 36000 samples vs OntoTune 360000 samples), making it easier to extend to other domains.

To demonstrate the generalizability of our method, we also conducted experiments using FIBO (Bennett, 2013) and Gene Ontology (Ashburner et al., 2000) as the ontology data sources for the finance and biology domains, respectively. We replicated previous ontology-based works, and the experimental results indicate that our method has stronger domain generalization.

Our contributions can be summarized as follows:

- We propose a novel fine-tuning framework based on ontology-driven prompts with innovative joint filtering strategy to construct thought chains for model domain reasoning alignment.
- Our approach significantly reduces data requirements for ontology knowledge, enabling efficient and low-cost adaptation across multiple domains.
- Compared to existing domain LLMs based on large-scale raw domain corpora and previous ontology-based approaches, our method achieves state-of-the-art performance in out-of-ontology domain QA tasks(i.e., MedMCQA(Pal et al., 2022), PubMedQA(Jin et al., 2019)), while effectively preserving the knowledge capabilities and safety of the seed model.

2 Related Work

2.1 Ontology enhanced LLMs

Traditional language model training typically relies on large-scale, unstructured text corpora, which lack explicit representations of conceptual semantic hierarchies (Wu et al., 2023). This limitation hinders the model's ability to acquire structured and systematic knowledge. To address this, recent research has increasingly explored the integration of hierarchical conceptual knowledge represented by ontologies (Staab and Studer, 2013) into language model training, aiming to enhance their semantic representation capabilities. Ontologies provide rich hierarchical structures and semantic constraints that help models better understand abstract relationships, organize domain-specific knowledge, and perform more effectively on tasks such as analogical reasoning, context disambiguation, and fine-grained classification. BioLORD-2023 (Remy et al., 2023) incorporates the UMLS ontology and adopts a combination of improved contrastive learning and self-distillation to strengthen semantic embeddings of biomedical concepts. Building on this, Ronzano and Nanavati (2024) leverage the MONDO ontology to further enhance biomedical semantic representations. Moskvoretskii et al. utilize WordNet (Miller, 1995) to fine-tune LLaMA (Touvron et al., 2023) under instruction supervision, resulting in TaxoLLaMA (Moskvoretskii et al., 2024), a vocabulary taxonomy model that achieves state-of-the-art results across four vocabulary classification benchmarks. Furthermore, Liu et al. (2025) introduce OntoTune, a self-training framework driven by the SNOMED CT medical ontology. By aligning LLMs with ontological structures through context-aware learning, OntoTune demonstrate strong domain-specific capabilities. These studies demonstrate the feasibility of using ontologies as a source of fine-tuning data.

In addition, some work focuses on the automated generation of ontologies, providing richer raw data for downstream data generation. OLLM (Lo et al., 2024) proposes an end-to-end ontology learning approach that fine-tunes a large language model on document-subgraph pairings to generate the target ontology. Nayyeri et al. (2025) proposed RIGOR, a framework that employs LLMs to automatically generate ontologies from relational databases.

2.2 CoT Distillation

Chain-of-thought (CoT) prompting has been shown to elicit intermediate reasoning processes from large language models and improve their performance on complex reasoning tasks (Wei et al., 2022; Kojima et al., 2022). Based on this observation, CoT distillation leverages rationales or

step-by-step reasoning traces as supervision signals, with the goal of transferring reasoning abilities to target models, either from stronger models or via self-distillation. Compared with answer-only supervision, CoT-style supervision provides richer intermediate learning signals and can improve both reasoning performance and data efficiency. For example, STaR (Zelikman et al., 2022) bootstraps reasoning traces from the model itself and iteratively refines the model through self-generated rationales. Distilling Step-by-Step (Kojima et al., 2022) shows that natural language rationales can help smaller models outperform larger models with fewer training examples. Orca (Mukherjee et al., 2023) further demonstrates that explanation traces from powerful teacher models can effectively improve the reasoning behavior of smaller student models. Recent studies continue this line of research by emphasizing the importance of reasoning-trace quality and selection. Quiet-STaR (Zelikman et al., 2024) extends self-training to latent rationales and shows that models can learn to generate useful intermediate thoughts before answering. s1 (Muennighoff et al., 2025) and LIMO (Ye et al., 2025) suggest that a relatively small amount of high-quality reasoning data can effectively elicit strong reasoning ability from pretrained models. These findings indicate that reasoning traces are not only useful supervision signals, but also efficient carriers for transferring structured reasoning behavior. Our work is inspired by CoT distillation and self-training, but differs from existing studies in both the knowledge source and the alignment objective. Prior work mainly focuses on general reasoning tasks or task-specific rationale construction. In contrast, OntoCoT uses ontology knowledge as a structured domain source and guides the seed model to generate associative CoT sequences.

3 Methodology

We will detail our method, as shown in Figure 1, in three stages:

- **Associative CoT Instruction Generation**: We use open-ended associative instruction to enable the base model to generate multiple thought chains under prompts with and without the corresponding ontology knowledge.
- **High-quality Instruction Text Selection**: We comprehensively consider both the similarity of the original text and the similarity of the thought chains. Using the model's self-scoring and clustering methods, we analyze the mastery level, knowledge richness, and the thought chain instructions to select high-quality thought chains as the training set.
- **LLM Fine-Tuning**: Based on the training set, we perform self-training on the seed model.

3.1 Associative CoT Instruction Generation

Considering the mind map-like structure present in the ontology data and the linear chain-like nature of human thinking, we aim to incorporate ontology relations into the thought chain process, so that the thinking process in the chain includes the additional knowledge provided by the ontology relationships.

Therefore, we use open-ended instructions to prompt the LLM to generate associative descriptions for the target ontology terms and the corresponding model's thought process. These descriptions are prompted to be organized as a chain-like structure using "→". For example, given the target ontology term *Periosteal bone formation, function (observable entity)*, a generated output is:

> Periosteal bone formation, function (observable entity) → Bone formation, function (observable entity) → Biological process of bone tissue development and repair → Response to mechanical stress, injury, or disease → Osteogenesis, osteoclastogenesis, and bone remodeling → Cellular and molecular mechanisms of bone development and maintenance → Hormonal regulation, growth factors, and mechanical forces → Bone development during fetal development and childhood growth → Developmental biology and pediatric medicine → Bone disorders and diseases, such as osteoporosis, bone fractures, and bone tumors → Clinical practice in orthopedics, traumatology, and endocrinology → Diagnostic and therapeutic approaches, including imaging techniques and surgical interventions → Understanding of periosteal bone formation, function in various physiological and pathological contexts.

This example shows the desired form of our training data: the chain starts from the target ontology term, incorporates its hierarchical context, and gradually expands toward biological mechanisms and clinically relevant applications. The specific corpus generation template is shown in Figure 5 of Appendix B. This template evaluates whether the model can relate specific ontology knowledge when encountering terms within the ontology. To mitigate the impact of randomness, we have the model generate N times. Denoting the open-ended instructions as x and the target ontology term that a single generation focuses on as t, the generation process is represented as:

$$y_t, y_c = g(\{f_\theta(x, t)\}_{i=1}^{N}) \quad (1)$$

where y_t is the list of the N original textual description generated by the model for the target

term, y_c is the list of associative CoT organized by the model, and the function g is a regular-based matching function used to extract the original textual description and thought chain from the model's generation.

To align the seed model with the ontology through contextual learning and enable the seed model to associate ontology information during the description generation process, we integrate ontology information related to the target term t into the input. The ontology information includes the definitions of the target term and the ontology structure of the target term (i.e., its hypernyms and synonyms), which are retrieved from the source ontology. The generation process with the concept's ontology information is represented as:

$$y_t^o, y_c^o = g(\{f_\theta(x, t, o_t)\}_{i=1}^N) \quad (2)$$

where o_t represents the ontology information related to the target term retrieved from the source ontology.

3.2 High-quality Instruction Text Selection

Dual-view Similarity Recall. For a target term t, if the model outputs y_t, y_c and y_t^o, y_c^o are inconsistent," it indicates that the model struggles to relate the corresponding ontology knowledge when reasoning about the target term. Therefore, we start with the model's reasoning text and perform selection based on its similarity. However, using text similarity often faces the problem of low sensitivity and is difficult to measure the differences in the thinking process within CoT data. As a result, we introduce CoT similarity, combining it with the text similarity of the original descriptions, and selecting the data with the lowest overall similarity from N generations for the target term t to obtain the data where the model's mastery is the weakest. For each original textual description in y_t^o , we calculate the embedding cosine similarity with each original textual description and take the average:

$$sim_{text_i} = \frac{1}{N}\sum_{j=1}^{N}\frac{E^T(y_{t_j})E(y_{t_i}^o)}{||E(y_{t_j})||||E(y_{t_i}^o)||} \quad (3)$$

where $y_{t_i}^o$ is an original text description from y_t^o, and $E(\cdot)$ is a sentence encoding model optimized for sentence-level similarity tasks (Reimers and Gurevych, 2019), which encodes the input sentence into a vector for semantic similarity evaluation.

To evaluate the similarity of reasoning chains, we design an evaluation metric. We believe that the importance of nodes located at different positions in a reasoning chain connected by "→" is different, and the difficulty with which the model can associate these nodes with the target term also varies. Nodes located closer to the front of the reasoning chain are more easily associated by the model. We hope that the introduction of ontological knowledge can help the model associate relevant knowledge from the ontology during the early stages of the generation process when facing questions involving the target term. This can act as a prompt to increase the likelihood of generating the correct answer. Therefore, we design a non-linear scoring mechanism that decreases from the head to the tail of the reasoning chain. The similarity of reasoning chains is then evaluated by the vector of scores assigned to different entities, as shown in Figure 2.

where we first split the reasoning chains in y_c and the corresponding list of y_c^o into lists of entities, then merge and deduplicate them. Afterward, we use the same encoding model $E(\cdot)$ to encode each entity in the list, and by comparing the pairwise cosine similarities and the predefined threshold $\delta \in [0, 1]$, we disambiguate the entities in the list, obtaining a dict of reasoning chain entities for the target term. Next, we apply our node scoring mechanism to obtain the scores for each entity in y_c, then merge and average these scores to construct a vector vec_{y_c} of the length of the dict. Afterward, we assign scores to the entities in a single CoT $y_{c_i}^o$ in y_c^o , resulting in the $vec_{y_{c_i}^o}$. Subsequently, we calculate the cosine similarity between each reasoning chain vector constructed based on y_c^o and vec_{y_c} as the similarity of the reasoning chain:

$$sim_{cot_i} = \frac{vec_{y_c} vec_{y_{c_i}^o}}{||vec_{y_c}||||vec_{y_{c_i}^o}||} \quad (4)$$

Finally, we add the original text similarity and the reasoning chain similarity to obtain the overall similarity score:

$$sim_i = sim_{text_i} + sim_{cot_i} \quad (5)$$

Self-evaluation Guided Selection. We select the reasoning chains with the lowest overall similarity in y_c^o as the initial training data $y_{c_i}^o$. However, the data obtained through dual-view similarity selection may still have issues with information redundancy and low quality. Therefore, we introduce clustering method and a self-evaluation module of the model for secondary filtering. First, we use UMAP (McInnes and Healy, 2018) to reduce the dimensionality of the sentence embeddings of the original text descriptions corresponding to $y_{c_i}^o$, and then apply the k-means (Kanungo et al., 2002) clustering algorithm to the data list composed of $y_{c_i}^o$.

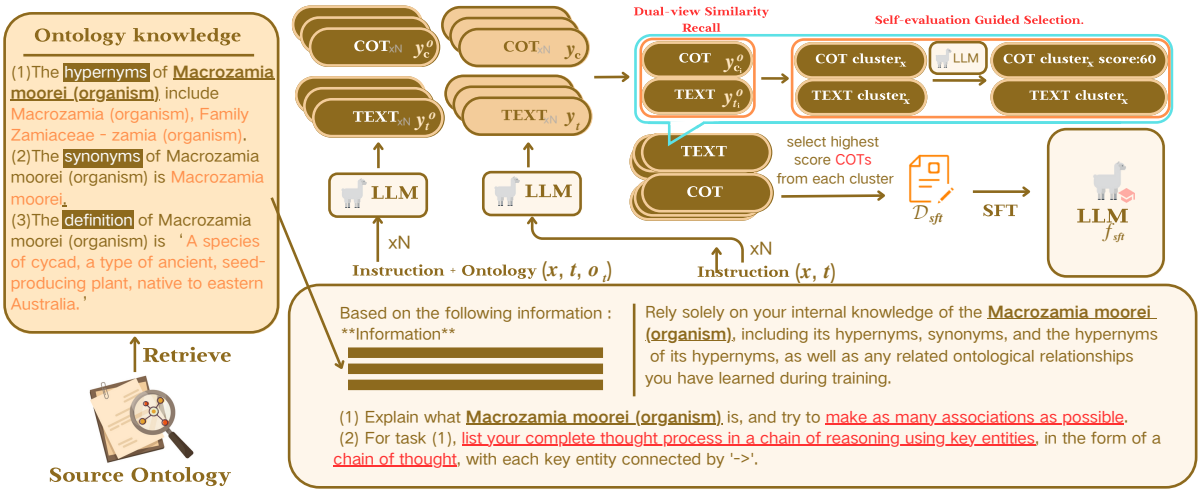

Figure 1: Overview of the pipeline of our method which construct training data based on ontologies.We show the detailed generation prompt and scoring prompt in Figure 5 of Appendix A.

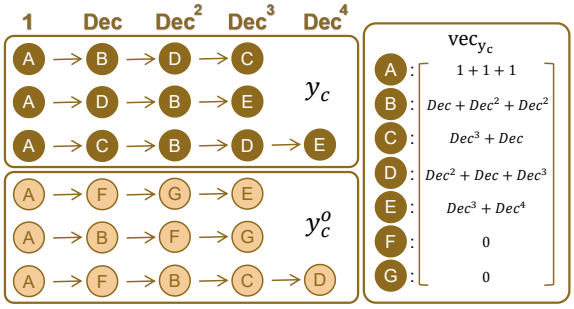

Figure 2: In this figure, the number of generations is denoted as $N = 3$. The entities represented by nodes such as A, B, etc., correspond to the entity nodes within the CoT. Given that the total number of distinct entities across all CoTs is 7, the dimensionality of the vectors is 7. The calculation method for the vector vec_{y_c} is outlined, where $Dec \in [0, 1]$ denotes the decay factor. The procedure for calculating the vectors for each CoT in y_c^o involves counting the entity scores only within the respective CoT.

We consider both the elbow method and the silhouette coefficient to select the final clusters.

Next, within each cluster, we use a scoring prompt based on various perspectives to evaluate the reasoning chains three times and compute the average score as the quality score. The scoring prompt is shown in Figure 5. Finally, we proportionally select the highest-scoring reasoning chains from each cluster based on the data distribution within the clusters.

3.3 LLM Fine-Tuning

Based on the data that we ultimately obtain, we modify the instruction to obtain the final training data $D = \{x'_n, y^o_n\}^k_{n=1}$ as shown in Figure 5. We then fine-tune the seed model using supervised instruction fine-tuning (SFT). Through SFT, we hope the seed model can learn the ontology-guided reasoning processes from y^o_t, improving its internal knowledge of the ontology. We use training data D to fine-tune the LLM f_θ directly with the next-token prediction objective for the response y^o_t:

$$\max_\theta \mathbb{E}_{(x'_t, y^o_t)\sim D} \left[\log P_\theta(y^o_t | x'_t)\right] \quad (6)$$

4 Experiment

In this section, we conduct a series of experiments to demonstrate the effectiveness and domain generalizability of our method. At the same time, to understand the impact of different hyperparameters and the role of various components in our method and try to make better use of the graph structure in ontology data, we also perform a series of analysis experiments. These experiments are designed to answer the following research questions:

- RQ1: Can our method adapt LLMs to specific domains, improving the performance of domain-specific tasks?
- RQ2: How does our method affect the general performance of the seed model?
- RQ3: Can our method be transferred across different domains and models?

4.1 Experiment setup

Following previous work, we select the medical domain as an example to evaluate the effectiveness of our method and extend it to the financial and biological domains to validate its stronger domain generality. In the medical domain, we use the standardized SNOMED CT (Donnelly, 2006) International Edition June version of 2024 as our source ontology, which includes 367,978 medical concepts. From this, we randomly sample 36,000 entries to practice our method, and after testing,

Setting	Model	MedQA	MedMCQA	PubMedQA	USMLE-step1	USMLE-step2	USMLE-step3	Average
zero-shot	LLaMA3 8B (Team, 2024)	51.7	51.7	70.3	57.4	52.3	58.2	56.9
	TaxoLLaMA (Moskvoretskii et al., 2024)	50.5	46.1	73.4	42.6	39.4	47.5	49.9
	OntoTune$_{sft}$ (Liu et al., 2025)	51.5	56.7	72.0	57.4	54.1	60.7	58.7
	OntoTune$_{dpo}$ (Liu et al., 2025)	**53.3**	**57.2**	65.5	58.5	51.4	59.0	57.5
	OntoTune$_{sft+dpo}$ (Liu et al., 2025)	51.9	56.7	66.3	53.2	54.1	**63.1**	57.6
	OntoCoT	52.1	56.8	**74.4**	**61.7**	**57.8**	59.8	**60.4**
SFT (on evaluation)	LLaMA3 8B (Team, 2024)	56.4	53.9	77.2	56.4	56	61.5	60.2
	Aloe (Gururajan et al., 2024)	53.4	56.8	75.4	54.3	**61.5**	60.7	60.4
	Med42-v2 (Christophe et al., 2024)	57.8	58.1	74.6	**60.6**	57.8	61.5	61.7
	TaxoLLaMA (Moskvoretskii et al., 2024)	55.9	57.5	77.6	56.4	57.8	59.0	60.7
	OntoTune$_{sft}$ (Liu et al., 2025)	58.4	60.4	78.6	57.4	57.8	62.3	62.5
	OntoTune$_{dpo}$ (Liu et al., 2025)	58.3	60.7	79.4	55.3	54.1	61.5	61.6
	OntoTune$_{sft+dpo}$ (Liu et al., 2025)	58.2	60.5	78.9	57.4	54.1	63.9	62.2
	OntoCoT	**59.1**	**61.3**	**79.6**	**61.7**	56.0	**68.0**	**64.3**

Table 1: Results of the medical domain QA in the zero-shot and supervised fine-tuning (on evaluation) setting. The best results are highlighted in bold, while the second best are underlined. Results are taken from OntoTune (Liu et al., 2025), and we evaluate our method in the same way.

we ultimately select $k = 9000$ high-quality samples for training. In the financial domain, we adopt the Q1 2025 version of FIBO (Bennett, 2013), extracting 2,329 concepts with ontology relationships (e.g., synonyms and superclasses) and definitions for data generation. In the biological domain, we use the basic version of Gene Ontology (Ashburner et al., 2000), extracting "hasExactSynonyms" as synonym relationships and "is_a" as superclass relationships, resulting in 44,737 concepts.

We use the LLaMA-3-8B-Instruct (Team, 2024) as our seed model due to its robustness and generalization across multiple domain tasks. We employ the Low Rank Adaptation(LoRA) (Hu et al., 2022) technique to fine-tune the model based on the LLaMA-Factory framework (Zheng et al., 2024). During the training phase of our method, we apply LoRA to all linear layers with a rank of $r = 8$. All training is conducted on 2 NVIDIA A100 80G GPUs. For the Supervised Fine-Tuning (SFT) stage, we use fp32 and a learning rate of 2×10^{-5}, training for 3 epochs with a cosine scheduler, a per-device batch size initialized to 8, and gradient accumulation of 2.

4.2 Medical Question Answering (RQ1)

To assess whether the seed model, after training with CoT data, can effectively generalize to domain-specific tasks, we perform an out-of-ontology domain QA task for evaluation.

Datasets. We utilize 6 medical QA datasets: MedMCQA (Pal et al., 2022), MedQA (Jin et al., 2020), PubMedQA (Jin et al., 2019), USMLE step1-3 (Han et al., 2023) datasets to comprehensively evaluate domain ability.

Baselines. To validate the effectiveness of our proposed method, we evaluate it against LLaMA3-8B-Instruct-based baselines: (1) established domain models built on large-scale corpora: Aloe (Gururajan et al., 2024) and Med42-v2 (Christophe et al., 2024); (2) approaches grounded in ontology: TaxoLLaMA (Moskvoretskii et al., 2024) and OntoTune (Liu et al., 2025). We perform instruction-tuning on the evaluation dataset's training set for our approach with the same training hyperparameters as OntoTune$_{sft}$ as the result in Table 1 under SFT(on evaluation).

Results. As shown in Table 1, OntoCoT yields substantial improvements over the original LLaMA3 8B baseline across both zero-shot and supervised settings. In the zero-shot setting, our method surpasses the vanilla model on all six benchmarks and achieves the highest average score among all compared methods, demonstrating strong robustness. Moreover, OntoCoT outperforms OntoTune variants on several key benchmarks. Under supervised fine-tuning, OntoCoT further attains the best overall average and ranks first on five of the six benchmarks, highlighting the effectiveness of our method. These findings indicate that structuring ontology knowledge in a CoT manner helps the model learn, organize, and exploit such knowledge more efficiently than direct ontology injection. To further examine whether the improvement comes from better domain reasoning rather than only answer memorization, we provide a qualitative case study in Appendix B.

4.3 General Capabilities and Safety Evaluation (RQ2)

Although our method uses a small amount of data, considering the potential overthinking caused by CoT-style data fine-tuning and the safety issues that may arise from open-ended thinking, we will evaluate whether the seed model retains its original general capabilities and whether it can maintain the

Model	MMLU					TriviaQA	Advbench	
	STEM	Social Sciences	Humanities	Other	Average	-	Raw Safe	Jailbreak Safe
Aloe	-1.16	-0.37	-1.90	-1.85	-1.39	-1.78	-35.00	-62.12
Med42-v2	-0.24	-0.37	-0.90	-1.43	-0.77	+0.38	-13.65	-36.16
TaxoLLaMA	-0.87	-2.87	-3.89	-4.67	-3.20	-1.69	-3.46	-23.08
OntoTune_sft	-0.36	-0.88	+1.04	-1.08	-0.18	-0.74	-3.46	-3.66
OntoTune_dpo	-0.50	-1.28	-0.88	-0.46	-0.79	-0.85	-6.92	-18.47
OntoTune_sft+dpo	-1.16	-1.44	+0.98	-1.39	-0.56	-0.85	-6.92	-11.54
OntoCot	-0.77	-0.37	-0.07	-2.72	-0.90	-1.63	-3.08	-5.20
LLaMA3 8B	56.83	76.61	60.81	74.10	66.49	64.81	97.50	96.35

Figure 3: Results of general capabilities and safety evaluation.

safety of its responses after applying our method.

Evaluation of General Capabilities. We conduct the general capabilities evaluation of our model on the MMLU (Hendrycks et al., 2020) and TriviaQA (Joshi et al., 2017). MMLU evaluation is performed using the LLaMA-factory (Zheng et al., 2024), while TriviaQA is evaluated using the OpenCampass (Contributors, 2023) tool with the gen mode. As shown in Table 3, our method performs almost on par with OntoTune in terms of general capabilities, without showing any significant performance degradation. Compared to the seed model, the average drop is only 1.08%, indicating that our approach does not introduce significant overthinking issues for the model.

Safety Evaluation. To verify whether open-ended thinking training data leads to safety issues in model responses, we use harmful instructions from the AdvBench (Zou et al., 2023) dataset as input and measure the proportion of safe responses from the model as the "Raw Safe" metric. Then, we test by adding an inducing suffix to the harmful instructions to prompt the model to give harmful responses, and we use the proportion of safe responses as the "Jailbreak Safe" metric.

As shown in the results in Figure 3, for the Raw Safe metric, the performance of our method is the best compared to other methods that fine-tune based on ontologies or large-scale domain-specific corpora. Under jailbreak settings, our method ranks just below OntoTune$_{sft}$, with no significant drop in safety performance. This indicates that the introduction of open-ended thinking training data still maintains safety alignment.

4.4 Domain and model Generalization Evaluation (RQ3)

To evaluate the domain adaptability of our method and its scalability with respect to ontology data size, we selected the finance and biology domains as our extension domains. We implemented our method on the ontology data from these domains and reproduced the method of OntoTune and TaxoLlama as baselines. To verify the scalability of our method on other models, we conducted an evaluation on LLaMA3.2-3B. All results reported in RQ3 are obtained under the zero-shot setting.

Model	MMLU_fi	CFA	Average
LLaMA3 8B (Team, 2024)	49.1	24.3	36.7
TaxoLLaMA (Moskvoretskii et al., 2024)	51.5	26.7	39.1
OntoTune$_{sft}$ (Liu et al., 2025)	44.1	23.6	33.9
OntoTune$_{dpo}$ (Liu et al., 2025)	49.6	24.6	37.1
OntoTune$_{sft+dpo}$ (Liu et al., 2025)	43.7	20.1	31.9
OntoCoT	**52.1**	**26.8**	**39.5**

Table 2: Results of financial Question Answering

Model	MMLU_bio	dbqa	Average
LLaMA3 8B (Team, 2024)	55.6	23.5	39.6
TaxoLLaMA (Moskvoretskii et al., 2024)	55.6	19.2	37.4
OntoTune$_{sft}$ (Liu et al., 2025)	54.2	17.9	36.1
OntoTune$_{dpo}$ (Liu et al., 2025)	55.6	18.3	37.0
OntoTune$_{sft+dpo}$ (Liu et al., 2025)	50.7	17.1	33.9
OntoCoT	**59.0**	**27.1**	**43.1**

Table 3: Results of biological Question Answering

Financial Question Answering. In the financial domain, we use FIBO (Bennett, 2013) as ontology source. Since the final number of selected ontology conceptions is less than 9000, we use the full dataset without clustering or intra-cluster scoring. Similarly, when reproducing the OntoTune and TaxoLLaMA methods, our data construction is also based on the full dataset. We choose CFA (Chartered Financial Analyst) and finance part of MMLU as our benchmarks, with evaluation instructions consistent with the medical domain.

OntoTune's three training strategies underperform the base model in Tables 2, likely due to difficulties adapting its instructions outside the medical domain. In contrast, OntoCoT achieves the best performance, demonstrating strong robustness and adaptability to small-scale data.

Biological Question Answering. In the biological domain, we use gene_ontology (Ashburner et al., 2000) as the source and filtering on the full dataset. For OntoTune, since the size of the full dataset is much smaller than its final selection, we directly use the full dataset for fine-tuning. To align with the ontology knowledge, we choose the dbqa subtask of the lab bench (Laurent et al., 2024) dataset and the biology part of MMLU as benchmarkswith evaluation instructions following the medical domain format. Similar to the finance

Model	MedQA	MedMCQA	PubMedQA	USMLE-step1	USMLE-step2	USMLE-step3	Average
LLaMA3.2 3B	**53.4**	52.5	65	56.4	78.0	78.7	64.0
OntoCoT	51.7	**53.2**	**65.5**	**62.8**	**82.6**	**82.0**	**66.3**

Table 4: Transfer Experiments on LLaMA3.2 3B.

Model	MedQA	MedMCQA	PubMedQA	USMLE-step1	USMLE-step2	USMLE-step3	Average
OntoCoT (N=6)	**52.1**	**56.8**	74.4	61.7	**57.8**	59.8	**60.4**
w/o self-score	52.0	56.0	74.3	**63.8**	52.3	**60.7**	59.9
w/o clustering	51.8	56.2	75.1	61.7	55.0	56.6	59.4
w/o CoT similarity	50.7	55.9	**75.2**	59.6	55.0	55.7	58.7
N=3	49.8	56.5	73.4	60.6	53.2	59.8	58.9

Table 5: Ablation studies of each component in our efficient selection strategy and the analysis studies on the number of generations(N).

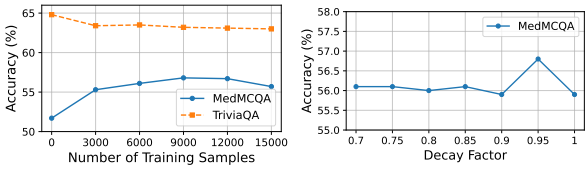

Figure 4: Performance with different number of training samples and decay factors.

domain, OntoCoT also achieves significant performance improvements, with a 3.6% increase on dbqa and a 3.4% increase on MMLU_bio (see Table 3), demonstrating OntoCoT's strong cross-domain extensibility.

Model Transfer Experiment. Due to the model size of LLaMA3.2-3B, we used a larger batch size and a smaller learning rate during fine-tuning. Table 4 shows that under zero-shot conditions, the average ability of llama3.2-3B outperforms llama3-8B, indicating that llama3.2-3B has stronger capabilities in the medical field. Our method significantly improves the performance of the seed model on five test sets (except Medqa) with an average improvement of 2.3%, demonstrating the model's transferability with our method.

4.5 Further Analysis

Ablation Study. To assess the contribution of each component in our selection strategy, we conduct ablation experiments by removing CoT similarity filtering, clustering, and self-scoring modules. All experiments fix the generation count $N = 6$, decay factor at 0.95, and use 9000 training samples. Results are shown in Table 5. Removing self-scoring (random selection within clusters) led to a relatively significant performance decrease on medmcqa and USMLE-step2. Omitting clustering results in less balanced outcomes across test sets, while discarding CoT similarity consistently harms performance, indicating its effectiveness in capturing fine-grained distinctions.

Hyperparameter Analysis. Table 5 shows that decreasing the number of generations (N), the performance of most test sets declines significantly, indicating that larger N effectively enhances the resulting training data quality. Figure 4 shows the performance of OntoCoT across different numbers of samples and decay factors. We use TriviaQA to evaluate general performance and MedMCQA to evaluate the domain-specific performance of the model with N fixed at 6. As the number of samples increases, OntoCoT gradually converges. When the number of samples reaches 9000, OntoCoT achieves the best domain-specific performance, indicating OntoCoT can achieve more robust results using significantly fewer training samples. A plausible explanation for the performance drop at 15,000 training samples is that, as the training set continues to expand, the proportion of less informative or lower-quality instances may increase because of our The limited complete data extracted from the raw data. At the same time, we find that there is an obvious performance peak when the decay factor is 0.95.

5 Conclusion

In this paper, we propose OntoCoT, an ontology-driven self-evolution framework that uses CoT as a knowledge carrier. OntoCoT focuses on model domain reasoning alignment and leverages the CoT containing ontology knowledge generated by the model itself for self-training, helping correct internal knowledge relationships and improve alignment between the seed model and the ontology. Experiments demonstrate that OntoCoT achieves state-of-the-art performance on ontology-independent medical QA tasks with relatively small dataset and show OntoCoT's strong model and domain generalization. We hope OntoCoT can inspire more methods that fully integrate graph-structured ontology knowledge into LLMs, enabling more effective ways to enhance model performance in specific domains with small-scale data.

Limitations

While OntoCoT demonstrates strong performance and generalization across multiple domains, there are several limitations to our current study:

- **Limited exploitation of multi-hop ontology knowledge.** In our current approach, we primarily leverage immediate ontology relations, such as direct parent terms and synonyms, to construct Chain-of-Thought sequences. The potential of more distant, multi-hop relationships within ontologies has not been fully explored. Investigating effective ways to incorporate such higher-order relations could further enhance reasoning and domain knowledge alignment.

- **Partial ontology usage due to computational and cost constraints.** For the medical domain, we used a subset of 36,000 SNOMED CT concepts instead of the full ontology, which contains over 360,000 concepts. This limited usage may restrict the coverage and completeness of the domain knowledge integrated into the model. Future work should explore the application of OntoCoT on the full ontology to evaluate its effectiveness at scale.

References

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, and 1 others. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Mike Bennett. 2013. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, 14(3):255–268.

Clément Christophe, Praveen K. Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *CoRR*, abs/2408.06142.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models.

Kevin Donnelly. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in Health Technology & Informatics*, 121:279.

Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrián Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, Sergio Álvarez-Napagao, Eduard Ayguadé Parra, Ulises Cortés, and Dario Garcia-Gasulla. 2024. Aloe: A family of fine-tuned open healthcare llms. *CoRR*, abs/2405.01886.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca - an open-source collection of medical conversational AI models and training data. *CoRR*, abs/2304.08247.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2567–2577. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):881–892.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard

Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 5848–5864. Association for Computational Linguistics.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. 2024. Lab-bench: Measuring capabilities of language models for biology research. *CoRR*, abs/2407.10362.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *CoRR*, abs/2307.10485.

Zhiqiang Liu, Chengtao Gan, Junjie Wang, Yichi Zhang, Zhongpu Bo, Mengshu Sun, Huajun Chen, and Wen Zhang. 2025. Ontotune: Ontology-driven self-training for aligning large language models. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 119–133. ACM.

Andy Lo, Albert Q. Jiang, Wenda Li, and Mateja Jamnik. 2024. End-to-end ontology learning with large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Leland McInnes and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024. Taxollama: Wordnet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2331–2350. Association for Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 20275–20321. Association for Computational Linguistics.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR*, abs/2306.02707.

Mojtaba Nayyeri, Athish A Yogi, Nadeen Fathallah, Ratan Bahadur Thapa, Hans-Michael Tautenhahn, Anton Schnurpel, and Steffen Staab. 2025. Retrieval-augmented generation of ontologies from relational databases. *CoRR*, abs/2506.01232.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, Proceedings of Machine Learning Research, pages 248–260. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

François Remy, Kris Demuynck, and Thomas Demeester. 2023. Biolord-2023: Semantic textual representations fusing LLM and clinical knowledge graph insights. *CoRR*, abs/2311.16075.

Francesco Ronzano and Jay Nanavati. 2024. Towards ontology-enhanced representation learning for large language models. *CoRR*, abs/2405.20527.

Steffen Staab and Rudi Studer. 2013. *Handbook on ontologies*. Springer Science & Business Media.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Weiqi Wu, Chengyue Jiang, Yong Jiang, Pengjun Xie, and Kewei Tu. 2023. Do plms know and understand ontological knowledge? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3080–3101. Association for Computational Linguistics.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387.

Eric Zelikman, Georges Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah D. Goodman. 2024. Quiet-star: Language models can teach themselves to think before speaking. *CoRR*, abs/2403.09629.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 400–410. Association for Computational Linguistics.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043.

A Prompt Templates

Figure 5 presents the prompt templates used in OntoCoT, including the generation prompt for constructing ontology-guided associative CoT sequences and the scoring prompt for model-based quality evaluation. The generation prompt encourages the model to organize ontology-related knowledge into a chain-like structure, while the scoring prompt evaluates the quality, informativeness, and reasoning coherence of generated CoT data.

B Qualitative Case Study

To further illustrate how OntoCoT improves domain reasoning, we provide a representative case from the medical QA evaluation. The seed model selects an incorrect treatment option, while OntoCoT identifies the correct diagnostic step.

Question. A 26-year-old woman with common variable immunodeficiency comes to the office because of a 2-month history of night sweats. She also has had a 10-kg (22-lb) weight loss during the past 3 months despite no change in appetite. She has a 15-year history of recurrent sinusitis and otitis media. Her only medication is intravenous immune globulin. She is 163 cm (5 ft 4 in) tall and weighs 54 kg (120 lb); BMI is 21 kg/m². Vital signs are within normal limits. Examination shows cervical and epitrochlear lymphadenopathy. Which of the following is the most appropriate next step in management?

Prompt for scoring

You are an expert in medical knowledge and data evaluation.Your task is to score a thought chain based on:
1. Completeness (25%): Does the chain cover a comprehensive range of concepts?
2. Coherence (30%): Are transitions logical and meaningful?
3. Relevance (20%): Does the chain focus on the medical topic specified in the instruction?
4. Depth (15%): Does the chain include specific medical concepts?
5. Novelty (10%): Does the chain provide unique connections?

Prompt for generation (with and without ontology knowledge)

You are an expert in the field of medicine.

Based on the following information:
(1)The hypernym of [Term] includes [hypernyms].
(2)The synonym of [Term] includes [synonyms].
(3)The definition of [Term] is [definition].

with ontology knowledge

Rely solely on your internal knowledge of the target medical term {term}, including its hypernyms, synonyms, and the hypernyms of its hypernyms, as well as any related ontological relationships you have learned during training.

without ontology knowledge

Complete the following tasks:
(1) Explain what medical term [Term] is, the answer should start with <explain> and end with </explain>, and try to make as many associations as possible without listing them separately."
(2) For task (1), list your complete thought process in a chain of reasoning using key entities, in the form of a thinking chain, with each key entity connected by '→'. The whole thinking chain should start with <reason> and end with </reason>. (e.g., <reason> Blood group antigen WES^b^ → Antigen on red blood cells → Blood group classification → Cromer blood group system → ISBT symbol CROM9 → Glycoproteins in antigen structure → Blood transfusion compatibility → Immune response to foreign antigens → Clinical practice recognition → Avoidance of transfusion complications </reason>).\n"

Note: (1) Only the answers to each task are required, no extra text should be provided. (2) Please generate the answer only once, do not repeat your response. (3) The chain of thought should be related to the answer for task 1 you gave, and each node should ideally appear only once, unless repetition is explicitly necessary for deepening reasoning or establishing new logical connections, avoiding repeated mentions of the same terms or subsequences!

Modified instruction in final training data

You are an expert in medicine. Combine the knowledge of the target medical term you have mastered, explain what medical term [Term] is, and try to make as many associations as possible without listing them separately.Represent your thought process in the form of a thinking chain like 'Blood group antigen WES^b^ → Antigen on red blood cells → Blood group classification → Cromer blood group system → ISBT symbol CROM9 → Glycoproteins in antigen structure → Blood transfusion compatibility → Immune response to foreign antigens → Clinical practice recognition → Avoidance of transfusion complications' where every node is a key entity, without generating extra text."

Figure 5: Prompt and instruction template used for ontology-guided CoT generation and quality scoring.

Options. (A) Lymph node biopsy; (B) Methylprednisolone therapy; (C) Penicillin therapy; (D) Plasmapheresis; (E) Trimethoprim-sulfamethoxazole therapy.

Ground Truth. (A) Lymph node biopsy.

Seed Model Response.

> A USMLE-style question!Let's break it down step by step: The patient has common variable immunodeficiency (CVID), a primary immunodeficiency disorder characterized by impaired antibody production. The patient has a 2-month history of night sweats and a 10-kg (22-lb) weight loss, which are concerning symptoms. She has a history of recurrent sinusitis and otitis media, which is consistent with her underlying immunodeficiency. The only medication she is taking is intravenous immune globulin, which is a replacement therapy for CVID.Given these symptoms and the patient's underlying condition, I would suspect that the patient may have an underlying infection, possibly opportunistic, which is causing her symptoms.Considering the options: (A) Lymph node biopsy: This may be necessary to determine the cause of the lymphadenopathy, but it's not the most appropriate next step in management. (B) Methylprednisolone therapy: This is not a specific treatment for the patient's symptoms, and it may not address the underlying cause of the night sweats and weight

loss. (C) Penicillin therapy: This is not a specific treatment for the patient's symptoms, and penicillin is not typically used to treat opportunistic infections. (D) Plasmapheresis: This is not a standard treatment for CVID or the patient's symptoms. (E) Trimethoprim-sulfamethoxazole therapy: This antibiotic combination is commonly used to treat opportunistic infections such as Pneumocystis jirovecii pneumonia (PJP), which is a common complication of CVID.Therefore, the most appropriate next step in management is:(E) Trimethoprim-sulfamethoxazole therapyThis answer is supported by the patient's symptoms, her underlying immunodeficiency, and the potential for opportunistic infections.

OntoCoT Response.

The correct answer is (A) Lymph node biopsy.The patient has a history of recurrent sinusitis and otitis media, which suggests a possible underlying immunodeficiency. The development of night sweats and weight loss over the past 3 months is concerning for an underlying infection or inflammatory process. The presence of cervical and epitrochlear lymphadenopathy suggests that the patient may have lymphadenitis or lymphoma.A lymph node biopsy would be the most appropriate next step in management to determine the cause of the lymphadenopathy and to guide further treatment. This could involve a fine-needle aspiration biopsy or an excisional biopsy, depending on the size and location of the lymph nodes.Methylprednisolone therapy (option B) may be used to treat inflammatory conditions, but it is not the most appropriate next step in this patient's case. Penicillin therapy (option C) would not be effective in treating lymphadenitis or lymphoma. Plasmapheresis (option D) is a treatment for autoimmune disorders, but it is not typically used to treat lymphadenitis or lymphoma. Trimethoprim-sulfamethoxazole therapy (option E) may be used to treat infections, but it is not the most appropriate next step in this patient's case without further evaluation and diagnosis.

Analysis. This case highlights a clear difference in clinical reasoning between the seed model and OntoCoT. Both models recognize that the patient has common variable immunodeficiency (CVID) and that her recurrent sinusitis and otitis media are consistent with impaired antibody production. However, the seed model anchors on the general association between immunodeficiency and opportunistic infection. As a result, it prematurely selects trimethoprim-sulfamethoxazole therapy, even though it briefly acknowledges that lymph node biopsy may help determine the cause of lymphadenopathy.In contrast, OntoCoT better preserves the diagnostic priority of the case. It connects the patient's night sweats, significant weight loss, and cervical and epitrochlear lymphadenopathy with the possibility of lymphoma or another lymphoproliferative process, which is clinically important in CVID. Instead of directly choosing empirical antimicrobial therapy, OntoCoT identifies lymph node biopsy as the appropriate next step to establish the underlying cause before treatment. This comparison suggests that ontology-guided CoT fine-tuning helps the model move beyond superficial disease-treatment associations and organize domain knowledge into a more clinically appropriate reasoning path.

C Multi-hop Data Analysis

To make better use of the graph structure in ontology data, we further explore whether introducing more distant hierarchical relations can improve model performance. Specifically, we introduce higher-level parent terms from the target ontology term and connect them with "→" as part of the prompt. We randomly select 3,000 ontology terms with multi-hop hierarchical relationships and compare this setting with the one-hop setting, where only the direct parent term of the target ontology term is introduced. We evaluate the model's domain ability using the same test set as in RQ1.

As shown in Table 6, introducing multi-hop hierarchical relationships leads to improvements on some benchmarks, such as MedMCQA and PubMedQA. However, the overall performance decreases compared with the one-hop setting. We speculate that introducing more distant ontology relations may also introduce additional noise, and that simply representing multi-hop relations as a linear chain may not be the optimal way to incorporate higher-order ontology structure.

Model	MedQA	MedMCQA	PubMedQA	USMLE-Avg	Average
One-hop	**51.8**	55.5	71.7	**57.1**	**58.4**
Multi-hop	51.1	**56.4**	**72.9**	55.1	57.6

Table 6: Effect of multi-hop hierarchical relationships on model performance.