Evontree: Ontology Rule-Guided Self-Evolution of Large Language Models

Anonymous ACL submission

Abstract

Although Large Language Models (LLMs) perform exceptionally well in general domains, the problem of hallucinations poses significant risks in specialized fields such as healthcare and law, where high interpretability is essential. Existing fine-tuning methods depend heavily on large-scale professional datasets, which are often hard to obtain due to the privacy regulations. Moreover, existing self-evolution methods are primarily designed for general domains, which may struggle to adapt to knowledge-intensive domains due to the lack of knowledge constraints. In this paper, we propose an ontology rule guided method Evontree to enable self-evolution of LLMs in low-resource specialized domains. Specifically, Evontree first extracts domain ontology knowledge from raw models, then detects knowledge inconsistencies using two core ontology rules, and finally reinforces gap knowledge into model via self-distilled fine-tuning. Extensive evaluations on medical QA benchmarks using Llama3-8B-Instruct and Med42-V2 demonstrate the effectiveness of Evontree, which outperforms both the base models and strong baselines, achieving up to a 3.7% improvement in accuracy. Detailed ablation studies further validate the robustness of our approach.

1 Introduction

Recently, large language models (LLMs) have demonstrated remarkable performance across a broad spectrum of domains, which is primarily attributed to extensive pre-training and subsequent fine-tuning on massive and specialized datasets. For instance, domain-specific models such as BioBERT (Lee et al., 2020) and SciBERT (Beltagy et al., 2019), have shown that scaling pre-training to specialized corpora is crucial for superior downstream outcomes. More recent adaptation methods such as Med42-v2 (Christophe et al., 2024), further underscore that instruction-tuning with billions of biomedical tokens remains a standard prerequisite for achieving optimal model performance.

However, acquiring such large-scale, high-quality domain data poses significant challenges in data-sensitive fields like healthcare and finance, due to stringent privacy regulations and high annotation costs. While existing domain fine-tuning methods depend heavily on external supervision, they often overlook the rich implicit ontology knowledge already internalized by LLMs during their pre-training phase (Wang et al., 2024). Recent research has reached a consensus that LLMs can be viewed as implicit knowledge bases. Yet, without explicit constraints, these models frequently suffer from hallucinations and knowledge inconsistencies when applied to knowledge-intensive tasks. To bridge this gap, we introduce Ontology, a structured knowledge representation utilized by domain experts to describe concepts, relations, and attributes. Considering the ability to maintain logical integrity, experts define ontology rules to automatically check for consistency within a knowledge base. In low-resource scenarios, these structured rules offer a high-density form of supervision that is more accessible than massive raw corpora.

In this paper, we propose Evontree, an ontology rule guided framework designed for the self-evolution of LLMs in specialized domains with minimal human supervision. Rather than traditional fine-tuning that introduces vast amounts of external data, Evontree focuses on refining the model's internal knowledge. Our approach treats the LLM as a source of "raw" ontology, using expert rules to detect and correct its internal inconsistencies. Specifically, our framework consists of three main steps: (1) Knowledge Extraction: We explicitly extract the implicit subclass and synonym relationships embedded within the LLMs. (2) Consistency Detection: We introduce two core ontology rules (as shown in Table 1) to identify logical inconsistencies within the extracted knowl-

ID	Premise	Conclusion
R1	$(x, \text{SynonymOf}, y) \wedge (y, \text{SubclassOf}, z)$	$\Rightarrow (x, \text{SubclassOf}, z)$
R2	$(x, \text{SubclassOf}, y) \wedge (y, \text{SubclassOf}, z)$	$\Rightarrow (x, \text{SubclassOf}, z)$

Table 1: Ontology Rules Used.

edge. (3) Knowledge Reinforcement: We re-inject the corrected and refined knowledge back into the model via self-distilled fine-tuning.

Experimental results demonstrate that Evontree consistently improves performance across major benchmarks. Notably, our method allows Llama3-8B-Instruct and the highly-tuned Med42-v2 to achieve significant accuracy gains (up to 3.7%), outperforming baselines that rely on large-scale supervised corpora. This confirms that leveraging a limited set of high-quality ontology rules can be more effective for domain adaptation than simply increasing data volume.

Our main contributions are as follows:

- We are the first to utilize ontology rules for low-resource LLM self-evolution. By leveraging llm's internal knowledge and simple ontology rules, we adapt LLMs to diverse professional fields where high-quality data is scarce.
- We propose a novel framework that integrates the extraction of implicit knowledge, rule-based refinement, and self-distilled re-injection, enabling effective self-evolution without external supervision.
- Experimental results demonstrate that Evontree outperforms post-training methods that rely on significantly larger supervised datasets, validating the effectiveness of our framework.

2 Preliminary

2.1 Ontology

Ontology captures concepts, their inter-connections, and rules within a specific domain. Specifically, (1) **Concepts** represent entities or categories within a domain, such as "Cell" in a medical ontology. (2) **Relationships** define how concepts are inter-connected. The most important relationships in ontologies are Hyponymy (Is-subclass-of) and Synonymy (Is-synonym-of). Hyponymy represents a hierarchical and subclass relationship. Synonymy indicates that two concepts are semantically equivalent. For example, "Muscle Cell" is a subclass of "Cell", while "Muscle Cell" and "Muscle Fiber" are synonyms. (3) **Axioms (Rules)**: Ontologies are equipped with built-in rules which enable automated reasoning and consistency checking within knowledge graphs. For example, if (Concept A, Is-subclass-of, Concept B) and (Concept B Is-subclass-of Concept C), then it logically follows that (Concept A, Is-subclass-of, Concept C). However, if the ontology also includes (Concept A, Is-Not-A-subclass-of, Concept C), a conflict arises with the inferred relationship. Such rules allow ontologies to automatically detect and resolve inconsistencies, ensuring the integrity of the knowledge graph. This capability is particularly valuable in large-scale knowledge bases, where manual verification is impractical.

2.2 Perplexity

Model's perplexity plays a critical role in our calculation of metric `ConfirmValue`. It quantifies the uncertainty of a probabilistic model in its predictions, where lower perplexity values correspond to higher prediction accuracy, while higher values indicate poorer performance. Formally, perplexity is defined as the exponential of the cross-entropy between the true distribution:

$$\text{Perplexity} = 2^{H(p,q)} \tag{1}$$

where $H(p, q)$ is the cross-entropy between the true distribution p and the model's predicted distribution q. In our framework, we leverage next-token prediction perplexity (Moskvoretskii et al., 2024; Li et al., 2024) to design our metric `ConfirmValue`, which evaluates model's confirmation towards certain ontology triple relationship. This helps mitigate hallucinations that arise from one-time generation.

3 Methodology

As illustrated in Figure 1, our method comprises the following key steps: (1) Ontology Knowledge Extraction, (2) Rule-Driven Ontology Examination, and (3) Gap Ontology Knowledge Injection.

3.1 Ontology Knowledge Extraction

Raw Triples Extraction: Firstly, we use 15 root concepts to initiate the ontology extraction process, including Antibiotic, Bacterium, Cell, Enzyme, Fungus, Hormone, Tissue, Vertebrate, Virus, Vitamin, Chemical, Inorganic Chemical, Organic Chemical, Infectious Disease and Non-Infectious Disease. In order to ensure a sufficient volume

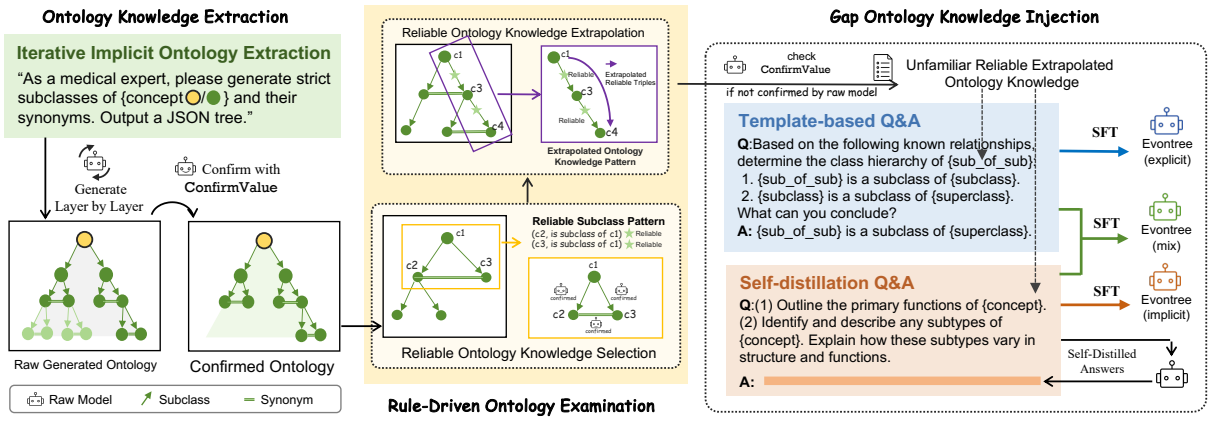

Figure 1: The overview of Evontree.

of gap ontology knowledge for fine-tuning , we restrict these root concepts' hierarchy depth≥ 3.

During the raw generation, for each concept c (initiated by root concept we manually designated, then iteratively replaced by generated last-layer concept), we feed the model the prompt as follows:

Generate Ontology Tree

As a medical expert, please generate strict subclasses of {concept} and their synonyms. Output a JSON tree like below:

```
{"{c}": {
    "description": "",
    "subclasses": [{
        "name": "",
        "description": "",
        "synonyms": ["", ""]
}]}}
```

After completing the iterative generation process, we obtain one ontology tree for each root concept, spanning at least three hierarchical layers. The initiated root concept is the apex of the ontology tree, and the underlying nodes are sub-concepts and synonym concepts generated layer-by-layer by the model itself. Edges of the tree denote the two most prevalent and fundamental ontology relations, subclass and synonym relationships.

`ConfirmValue` **Computation:** To reduce hallucination from one-time generation of models, we design the metrics `ConfirmValue`, a causal-perplexity(introduced in Section 2.2)-based confidence score, to evaluate model's confirmation extent towards a single ontology triple. Taking raw triple $(A, \text{SynonymOf}, B)$ as an example, we probe the model with the prompt as following.

Calculate ConfirmValue

Please determine if the statement is true or false, then answer with True or False: 'A' is an exact synonym of 'B'. Answer:

For each triple t under prompt p, we construct two completions:(1) $S_{\text{True}} = p +$ "True", (2) $S_{\text{False}} = p +$ "False", and compute the perplexity over the "Answer: True/False" token span. The `ConfirmValue` is defined as

$$\texttt{ConfirmValue}(t, p) = \frac{\text{sign}\big(\text{PPL}_{\text{False}} - \text{PPL}_{\text{True}}\big)}{\min\big(\text{PPL}_{\text{True}},\ \text{PPL}_{\text{False}}\big)}. \tag{2}$$

To robustly gauge model's confirmation towards certain ontology triple, we craft five paraphrased prompts for synonym relations and four for subclass relations. A triple is deemed model-confirmed only when every associated `ConfirmValue` surpasses a threshold τ.

Confirmation Threshold Setting: The threshold τ is obtained for each model and each prompt. For each model–prompt pair, we regard the one-time generated output (true / false) as a binary decision signal and `ConfirmValue` as its continuous confidence score. Using the ≈ 10 k raw subclass and synonym ontology triples from raw generation, we sweep $\tau \in [0, 1]$ and compute the true-positive and false-positive rates. The optimal threshold τ^* is the cut-off that maximizes the Youden index $J = \text{TPR} - \text{FPR}$, i.e. the point where the binary decision and continuous confidence are best aligned. If the model designates certain ontology triple with `ConfirmValue` higher than threshold τ, then we deem that this ontology triple is "confirmed" by the raw model. If the model assigns a `ConfirmValue` exceeding the threshold τ to a given ontology triple, we regard that triple as "con-

firmed" by the raw model.

3.2 Rule-Driven Ontology Examination

Reliable Ontology Knowledge Selection: While the preceding steps yield triples that are confirmed by the model, such confirmation alone does not ensure factual reliability. To guarantee that only robust knowledge is employed for downstream extrapolation, we must avoid arbitrarily selecting triples that may be erroneous. Therefore, we apply ontology rule R1 to identify closed triangles consisting of two subclass triples and one synonym triple, where the mutual structure allows each edge to corroborate the others. Any subclass triple that forms part of such a triangle is considered reliable. This rule leverages both ontological constraints and our prior belief that mislabelled instances within the model's pre-training corpus are rare; therefore, the likelihood of three mutually supporting but erroneous statements co-occurring is extremely low. Consequently, only reliable subclass triples are retained for extrapolation in the subsequent step.

Ontology Knowledge Extrapolation: We take the reliable subclass triples $\mathcal{T}_{\text{rel}}$ from last step and apply ontology rule R2 to generate new extrapolated subclass triples $\mathcal{T}_{\text{extrapolated}}$. For example, given reliable triples $(D,\ \text{subClassOf},\ C)$ and $(C,\ \text{subClassOf},\ A)$, rule R1 yields $(D,\ \text{subClassOf},\ A)$.

Gap Ontology Triples Selection: For every $t_{\text{extrapolated}} \in \mathcal{T}_{\text{extrapolated}}$, we recompute its `ConfirmValue` using the same prompt set as the raw ontology extraction step. Crucially, we retain only those triples whose `ConfirmValue` is below the threshold τ^* , which means those triples the model are not familiar with before. These low-ConfirmValue triples are labelled $\mathcal{T}_{\text{gap}}$, which are supposed to be injected into the model during the subsequent model's fine-tuning stage.

3.3 Gap Ontology Knowledge Injection

Explicit Injection: We introduce three injecting strategies for incorporating these credible but previously unfamiliar triples into the model. The simplest and most straightforward approach is to leverage our reliable and extrapolated triples to construct explicit reasoning chains, which are then used to synthesize question–answer pairs; the generation template is illustrated in the main figure 1.

Implicit Injection: To mitigate homogeneity in synthesized QA data from single-prompt reasoning chains, we follow prior work (Liu et al., 2025) to generate more natural and diverse training corpus. We append the ontology chain to pre-defined, concept-specific question templates in order to guide model to produce concept-aware higher-quality answers. We then finetune the model with the instruction and self-distilled output pairs back into the model, ensuring the ontology knowledge and concept-aware, self-distilled knowledge is reinforced to the model to continually enhance model's domain capability.

The two pre-define concept-centric question templates are shown in below.

Generate Corpus

(1) Outline the primary functions of {concept}.
(2) Identify and describe any subtypes of {concept}. Explain how these subtypes vary in structure and function.

For each gap triple $(D,\ \text{SubclassOf},\ A)$ and its derivation chain $(D,\ \text{SubclassOf},\ C)$, $(C,\ \text{SubclassOf},\ A)$, we instantiate the placeholders with the three involved concepts A, C, D for template (1), and with A, C for template (2).

We hypothesise that appending the derivation chain as a *hint* after each template enables the model to produce higher-quality answers. The hint template are shown in following substantiated example. Given the weak triple (Skeletal Muscle Fiber, SubclassOf, Cell) supported by two reliable chains:(i) (Skeletal Muscle Fiber, SubclassOf, Muscle Cell), (Muscle Cell, SubclassOf, Cell). (ii) (Skeletal Muscle Fiber, SubclassOf, Myocyte), (Myocyte, SubclassOf, Cell). we generate the following instruction for chain (ii):

Hint with Ontology Chain

Outline the primary functions of Skeletal Muscle Fiber. You can consider these relationships as follows, but please ignore them if they are unnecessary: Skeletal Muscle Fiber is a subclass of Myocyte, and Myocyte is a subclass of Cell.

The same process is repeated for every supporting chain.

Mixed Injection As a further variant, we simultaneously inject knowledge via both implicit and explicit ways, which is termed mixed injection.

4 Experiments

We designed four research questions to verify the effectiveness of our method: (RQ1) Can our method accurately identify factually reliable ontology knowledge that the raw model fails to master? (RQ2) Can our method effectively improve

raw model's domain capability without introducing any external corpora? (RQ3) How does Evontree impact raw model's generalization performance? (RQ4) Does every key component of Evontree make a meaningful contribution?

4.1 Experiment Settings

4.1.1 Datasets

Following prior work (Liu et al., 2025; Gururajan et al., 2024; Christophe et al., 2024), we adopt 3 widely-used medical datasets, including PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) for evaluation.

4.1.2 Baselines

We use Llama-3-8B-Instruct as our base model and additionally include Med42-v2(Christophe et al., 2024), a Llama-3-8B variant fine-tuned in medical domains to test the effectiveness of our framework on an already strong domain model.

4.1.3 Implementation

Models are fine-tuned using LoRA via LLaMA-Factory(Zheng et al., 2024) with a learning rate of $5e-5$ for 3 epochs. During the SFT stage, we use fp32 and a learning rate of 5e-5, training for 3 epochs with a cosine scheduler, a batch size per device initialized to 8 and gradient accumulation of 2.

4.1.4 Evaluation

We evaluate the effectiveness of our approach from two perspectives. On the one hand, we employ GPT-4o-mini(OpenAI, 2023) and DeepSeek-V3(DeepSeek-AI et al., 2024) as strong supervisory models to assess the quality of the ontology triples generated at each stage of our pipeline. On the other hand, we measure our model's performance on both medical-domain QA datasets and general capability benchmarks, following the evaluation protocols established in prior work (Liu et al., 2025).

4.2 Characteristics of Ontology Triples (RQ1)

We observe the characteristics of ontology knowledge at different stages by analyzing the relationship between their factual accuracy and the model's self-confidence (`ConfirmValue`), as illustrated in Figure 2.

Effectiveness of Reliable Ontology Knowledge Selection. A direct comparison between *Confirmed triples* (blue) and *Reliable triples* (green) validates

Triple Type	Relation	Num	`ConfirmValue` Avg.	Acc.
Raw	synonym	95,814	0.0537	54.90
Confirmed	synonym	30,532	0.4738	87.51
Raw	subclass	58,309	0.2094	47.79
Confirmed	subclass	24,276	0.4896	72.62
Reliable	subclass	14,169	0.4981	74.81
Extrapolated	subclass	8,661	0.4449	80.29
Gap	subclass	1,396	0.1595	74.57

Table 2: Ontology path statistics and quality evaluation.

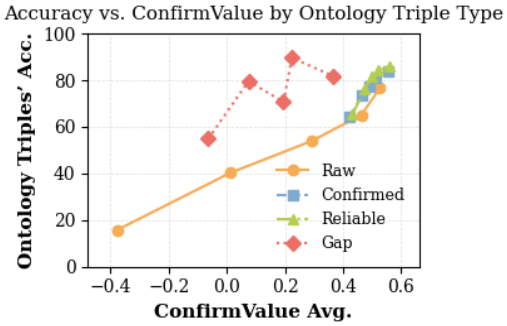

Figure 2: Relationship between accuracy and `ConfirmValue`.

the effectiveness of our **Reliable Ontology Knowledge Selection** module. While both types occupy a similar range of high `ConfirmValue`, the reliable triples consistently achieve higher accuracy across the spectrum. This demonstrates that our unsupervised selection strategy successfully identifies more factually accurate knowledge beyond what simple confidence thresholds can provide.

Identifying Knowledge Gaps. The contrast between *Gap triples* (red) and the other categories (blue and green) reveals a critical finding. Although the accuracy of gap triples is comparable to confirmed and reliable ones, their average `ConfirmValue` is significantly lower. This pattern aligns perfectly with our core hypothesis: Evontree effectively isolates "blind spots"—knowledge that is factually correct but for which the raw model lacks internal certainty. By targeting these low-confidence but high-accuracy triples, we ensure that the self-evolution process focuses on the most valuable and reliable information gaps.

4.3 Evaluation on Medical Datasets (RQ2)

As shown in Table 3, among the three Evontree variants, the "explicit" ontology injection approach does not effectively improve model performance, whereas the more natural "implicit" and "mix" variants yield significant enhancements. This trend is further validated when comparing

TaxoLLaMA (which uses explicit ontology injection) and OntoTune (which adopts implicit injection). Notably, Evontree not only achieves substantial performance gains for general-purpose LLMs (e.g., LLaMA3-8B-Instruct), but also further boosts domain-specific models like Med42-v2, reaching improvements of 3.1% and 3.7%, respectively. Compared with existing LLaMA3-8B-based domain models, Evontree attains state-of-the-art results even without introducing additional external supervision data, outperforming baselines by an average of 1.1%. This indicates that large-scale, fragmented corpora alone cannot reliably achieve effective domain alignment, and may even introduce conflicting knowledge that increases model confusion. We conclude that the advantage of Evontree lies in its targeted identification of effective gap triples within the model's knowledge system. Through fine-grained automated selection, Evontree ensures that only information highly compatible with the model's existing knowledge structure is injected. This mechanism enables the model to self-reflect and precisely locate knowledge gaps, evolving itself based on reliable information while effectively avoiding unnecessary noise and conflicts. Compared to traditional data augmentation and domain adaptation approaches, Evontree prioritizes quality over quantity in knowledge supplementation, promoting greater consistency and completeness in the model's knowledge base.

4.4 General Capabilities and Safety Evaluation (RQ3)

General Capabilities Evaluation. We evaluate the general capabilities of our model variants compared to the raw model on the MMLU, TriviaQA, and ARC datasets. Specifically, MMLU is assessed using the LLaMAFactory framework, while ARC and TriviaQA are evaluated using the OpenCompass tool in gen mode. As shown in Table 4, under the zero-shot setting, all three variants of our method exhibit no significant degradation in general capabilities. The best-performing variant shows an average performance drop of only 0.15% compared to the raw model, and LLaMA3-Evontree(mix) even achieves marginal improvements on certain test sets.Furthermore, under the supervised fine-tuning (SFT) evaluation setting, our models continue to demonstrate similar trends. Compared to med42-llama3-8b, our variants maintain stable performance and in some cases achieve notable gains, highlighting the robustness and potential benefits of our approach.

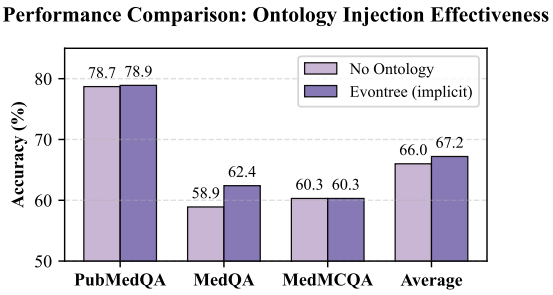

Figure 3: Ablation Study of Med42-v2.

Safety Evaluation. Following prior work, we also assess whether our method introduces any safety risks to the model. We use harmful instructions from the AdvBench dataset to measure the proportion of safe responses, reported as the "Raw Safe" metric. Then, we append an inducing suffix to the harmful prompts to encourage unsafe behavior and measure the proportion of safe outputs under this adversarial setting, reported as the "Jailbreak Safe" metric. As shown in Table 4, the Evontree(implicit) and Evontree(mix) variants show no significant drop in safety compared to the raw model, and in fact, when built upon Med42, these two variants even lead to notable improvements. In contrast, the Evontree(explicit) variant performs worse in both the LLaMA3 and Med42 settings. We hypothesize that, similar to the observations in TaxoLLaMA, the introduction of rigid template-based Q&A formats may negatively impact the model's safety performance.

4.5 Ablation Study (RQ4)

The ablation results in Table 6 and Figure 3 demonstrate that Evontree's performance stems from the synergy of its three core modules: (1)**Ablation on Reliable Ontology Knowledge Selection.** The significant accuracy gap in Table 5 (74.6% vs. 55.7%) underscores that simple confidence scores (`ConfirmValue`) are insufficient for identifying truth, justifying our rule-based reliable selection. (2) **Ablation on Gap Triple Selection.** The performance drop across all three medical benchmarks when removing Gap Triple Selection (Table 6) confirms that the model benefits most from knowledge it is "uncertain" about, rather than redundant information. (3)**Ablation on Ontology Injecting.** The fact that the "w/o ontology injecting" variant declined the most indicates that structured ontology serves as a vital anchor for reasoning, preventing

Setting	Model	MedQA	MedMCQA	PubMedQA	Average
zero-shot	LLaMA3 8B (Dubey et al., 2024)	51.7	51.7	70.3	57.9
	TaxoLLaMA (Moskvoretskii et al., 2024)	50.5	46.1	73.4	56.7
	OntoTune$_{sft}$ (Liu et al., 2025)	51.5	56.7	72.0	60.1
	OntoTune$_{dpo}$ (Liu et al., 2025)	**53.3**	57.2	65.5	58.7
	OntoTune$_{sft+dpo}$ (Liu et al., 2025)	51.9	56.7	66.3	58.3
	LLaMA3 8B-Evontree (explicit)	43.4	48.6	**76.4**	56.1
	LLaMA3 8B-Evontree (implicit)	52.7	54.4	72.9	60.0
	LLaMA3 8B-Evontree (mix)	51.0	**57.4**	74.7	**61.0**
	Δ Improvement over raw model	+1.0%	+5.7%	+6.1%	+3.1%
	Δ Improvement over best baseline	-0.6%	+0.2%	+3.0%	+0.9%
SFT (on evaluation)	LLaMA3 8B (Dubey et al., 2024)	56.4	53.9	77.2	62.5
	Aloe (Gururajan et al., 2024)	51.1	56.8	75.4	61.1
	Med42-v2 (Christophe et al., 2024)	57.8	58.1	74.6	63.5
	TaxoLLaMA (Moskvoretskii et al., 2024)	55.9	57.5	77.6	63.7
	OntoTune$_{sft}$ (Liu et al., 2025)	58.4	60.4	78.6	65.8
	OntoTune$_{dpo}$ (Liu et al., 2025)	58.3	60.7	**79.4**	66.1
	OntoTune$_{sft+dpo}$ (Liu et al., 2025)	58.2	60.5	78.9	65.9
	Med42-v2-Evontree (explicit)	57.1	58.6	71.1	62.3
	Med42-v2-Evontree (implicit)	**60.3**	**62.4**	78.9	**67.2**
	Med42-v2-Evontree (mix)	57.2	57.7	74.9	63.3
	Δ Improvement over raw model	+2.5%	+4.3%	+4.3%	+3.7%
	Δ Improvement over best baseline	+1.9%	+1.7%	-0.5%	+1.1%

Table 3: 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.

the model from reverting to hallucinated patterns during self-distillation.

Collectively, these ablation studies confirm that Evontree's effectiveness stems from its ability to transform implicit, unverified model knowledge into explicit, factually-grounded domain expertise.

5 Related Work

5.1 Domain Adaptation and Knowledge Integration

Large Language Models (LLMs) often require specialized adaptation to perform effectively in knowledge-intensive fields (Xie et al., 2024). Early efforts, such as BioBERT (Lee et al., 2020) and BloombergGPT (Wu et al., 2023), relied on extensive pre-training on domain-specific corpora. More recent approaches like Med42-v2 (Christophe et al., 2024) and Aloe (Gururajan et al., 2024) focus on instruction-tuning using billions of tokens. However, as noted in recent studies, acquiring such large-scale professional datasets is often hindered by privacy regulations and high annotation costs in sectors like healthcare and finance.

To mitigate hallucinations and improve reasoning, ontology-integrated methods have emerged to align LLMs with structured knowledge. Frameworks such as TaxoLLaMA (Moskvoretskii et al., 2024) and OntoTune (Liu et al., 2025) incorporate external, human-curated ontologies (e.g., SNOMED CT) to provide semantic constraints.

In contrast to these data-dependent paradigms, our approach recognizes that LLMs already function as implicit knowledge bases (Wang et al., 2024). Rather than injecting external supervision, we focus on extracting and refining the model's internal ontology knowledge, achieving robust domain adaptation in low-resource scenarios without relying on any external annotated corpora.

5.2 Self-Evolution of LLMs

Self-evolution allows LLMs to autonomously generate training data, reducing the need for human labor. For instance, SELF-INSTRUCT (Wang et al., 2023b) enables models to bootstrap task instructions and responses. To ensure the quality of self-generated content, various selection heuristics prioritize logical consistency (Huang et al., 2023;

Model	MMLU					ARC		TriviaQA	Advbench	
	STEM	Social Sciences	Humanities	Other	Average	ARC_C	ARC_E	-	Raw Safe	Jailbreak Safe
LLaMA3 8B (Dubey et al., 2024)	56.83	76.61	60.81	74.10	66.49	78.64	92.77	64.81	97.50	96.35
TaxoLLaMA (Moskvoretskii et al., 2024)	55.96	73.74	56.92	69.43	63.29	72.88	89.24	63.12	94.04	73.27
OntoTune$_{sft}$ (Liu et al., 2025)	56.47	75.73	**61.85**	73.02	**66.31**	78.31	91.89	**64.07**	94.04	92.69
OntoTune$_{dpo}$ (Liu et al., 2025)	56.33	75.33	59.93	**73.64**	65.70	78.98	**92.06**	63.96	90.58	77.88
OntoTune$_{sft+dpo}$ (Liu et al., 2025)	55.67	75.17	61.79	72.71	65.93	78.98	**92.06**	63.96	90.58	84.81
LLaMA3-Evontree (implicit)	56.89	76.41	61.38	71.99	66.16	79.66	89.59	63.00	95.38	**96.15**
LLaMA3-Evontree (mix)	**56.99**	**76.80**	61.62	71.71	66.28	80.34	90.30	63.22	95.19	**96.15**
LLaMA3-Evontree (explicit)	56.89	76.60	60.85	72.24	66.08	**82.03**	91.71	63.42	**97.50**	83.65
Aloe (Gururajan et al., 2024)	55.67	76.24	58.91	72.25	65.10	75.25	86.95	63.03	62.50	34.23
Med42-v2 (Christophe et al., 2024)	56.59	76.24	59.91	72.67	65.72	**82.37**	**92.59**	**65.19**	83.85	60.19
Med42-v2-Evontree (implicit)	**56.86**	**76.37**	59.79	72.79	65.80	80.34	91.36	63.54	86.54	60.77
Med42-v2-Evontree (mix)	56.46	76.24	**60.09**	72.86	65.80	80.68	92.06	63.71	**89.23**	**65.77**
Med42-v2-Evontree (explicit)	56.79	76.11	59.98	**72.92**	**65.82**	80.68	91.18	64.33	75.38	45.19

Table 4: Results of general capabilities and safety evaluation.The results with the least decrease compared to the original model are highlighted in bold, while the second least are underlined.

Variant	gpt	deepseek	Avg.
w/o reliable triple sel.	58.4	53.0	55.7
Evontree	75.3	73.8	74.6

Table 5: Impact of Reliable-Triple Selection on Extrapolated Triple Accuracy.

Variant	MedQA	MedMCQA	PubMedQA	Avg.
w/o reliable triple sel.	52.4	**54.7**	72.3	59.8
w/o gap triple sel.	52.6	54.1	72.5	59.7
w/o ontology injecting	51.1	53.9	**73.1**	59.4
LLaMA3-Evontree(implicit)	**52.7**	54.4	72.9	**60.0**

Table 6: Ablation Study of Different Modules.

Wang et al., 2023a) or adhere to predefined ethical criteria (Sun et al., 2023). Additionally, Chain-of-Thought (CoT) distillation (Yu et al., 2024) is frequently used to enhance the reasoning depth of generated outputs.

However, existing self-evolution methods are primarily designed for general domains and often struggle in specialized fields due to a lack of rigorous knowledge constraints. Unlike these general-purpose heuristics, our framework, **Evontree**, introduces knowledge-related ontology rules as a supervisionenabling effective self-evolution in specilized domains.

6 Conclusion

In this work, we address the challenge of adapting large language models to data-scarce domains by leveraging the unique value of domain ontology rules. Our framework enables explicit extraction and validation of LLMs' implicit knowledge, using only two carefully chosen ontology rules to systematically detect and correct inconsistencies. By re-injecting revised knowledge into the model, we substantially improve performance on medical QA tasks, outperforming baselines dependent on large supervised datasets. Extensive experiments across two representative models and several medical benchmarks validate the robustness and effectiveness of our paradigm. This study highlights the potential of ontology rule-driven supervision as a practical and powerful solution for enhancing LLMs in highly specialized domains where conventional data-centric approaches are limited by privacy or scarcity. Future work can explore broader application of our framework to other professional domains and further enrichment of ontology rule-guided knowledge editing techniques.

7 Limitations

Despite its effectiveness, Evontree has several limitations that offer avenues for future work.

First, the framework currently requires 15 manually selected root concepts to initiate the ontology generation tree. While this is the only external input beyond our two rules, future research could develop a more automated "from-scratch" approach by leveraging self-evaluation and model consistency to identify adequate root concepts without human intervention.

Second, our experiments primarily focused on subclass and synonym relationships. While these are fundamental to most domain ontologies, the logic of Evontree is naturally extensible. Future work will explore more complex relationship types and diverse ontology rules, aiming to adapt our self-evolution framework to a broader range of specialized knowledge-intensive domains.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In

EMNLP/IJCNLP (1), pages 3613–3618. Association for Computational Linguistics.

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

DeepSeek-AI and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*. Assuming V3 release, otherwise use V2 citation: arXiv:2406.11931.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, and et al. 2024. Aloe: A family of fine-tuned open healthcare llms. *arXiv preprint arXiv:2405.01886*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *EMNLP*, pages 1051–1068. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, page 6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, WilliamW. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Cornell University - arXiv,Cornell University - arXiv*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *NAACL-HLT*, pages 7602–7635. Association for Computational Linguistics.

Zhiqiang Liu, Chengtao Gan, Junjie Wang, and et al. 2025. Ontotune: Ontology-driven self-training for aligning large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 119–133.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024. Taxollama: Wordnet-based model for solving multiple lexical sematic tasks. *CoRR*, abs/2403.09207.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*, pages 248–260. PMLR.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *NeurIPS*.

Mengru Wang, Yunzhi Yao, Ziwen Xu, Shuofei Qiao, Shumin Deng, Peng Wang, Xiang Chen, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. Knowledge mechanisms in large language models: A survey and perspective. In *EMNLP (Findings)*, pages 7097–7135. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In *ICLR*. OpenReview.net.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *ACL (1)*, pages 13484–13508. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. In *ACL (Findings)*, pages 10184–10201. Association for Computational Linguistics.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. 2024. Distilling system 2 into system 1. *CoRR*, abs/2407.06023.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.