Domain Knowledge Annealing: A Structured Local-to-Global Paradigm for Constructing Expert LLMs

Yikang Zhang[1,2], Zuchao Li[2]*, Ping Wang[3]*, Lefei Zhang[4], Hai Zhao[5], Bo Du[4]

[1]Academy of Advanced Interdisciplinary Studies, [2]School of Artificial Intelligence, [3]School of Information Management, [4]School of Computer Science, Wuhan University
[5]School of Computer Science, Shanghai Jiao Tong University
{zhangyikang,zcli-charlie,wangping,zhanglefei,dubo}@whu.edu.cn
zhaohai@cs.sjtu.edu.cn

Abstract

Expertise in vertical domains requires more than exposing LLMs to vast corpora; it demands the systematic learning of complex logic. Standard fine-tuning often treats data as an unstructured collection, failing to capture the dependencies needed for high-level reasoning. We propose Domain Knowledge Annealing (DKA), a curriculum learning paradigm designed to improve domain proficiency. Inspired by physical annealing, DKA organizes training from local to global: it starts by injecting discrete concepts via entity-centric samples and progressively advances to synthesizing complex, inter-document relationships. This progression allows the model to integrate isolated facts into a coherent knowledge system. Experiments on Linguistics and Law benchmarks show that DKA outperforms standard strategies, achieving state-of-the-art results among comparably sized models in domain-specific reasoning.

1 Introduction

The transformative potential of LLMs in vertical domains, such as medicine (Thirunavukarasu et al., 2023), finance (Li et al., 2023), and law (Sun, 2023), depends on their ability to function as rigorous experts. Unlike general chat scenarios where fluency suffices (Zhao et al., 2023; Minaee et al., 2024; Li et al., 2025), specialized fields require strict adherence to logical frameworks and synthesizing definitive conclusions from complex evidence. However, generic models frequently struggle with deep reasoning, exhibiting hallucinations and inconsistencies (Ge et al., 2023; Zheng et al., 2025), limiting deployment in high-stakes environments.

We attribute this to a misalignment between training dynamics and domain knowledge structure. True expertise resembles a structured network of interconnected concepts. Conversely, standard fine-tuning treats data as unstructured "bags of tokens," ignoring logical complexity. While recent work suggests "Textbooks Are All You Need" (Gunasekar et al., 2023), real-world domain data often remains fragmented. Traditional data synthesis exacerbates this by extracting isolated facts from intra-document contexts, yielding simplistic question-answer (QA) pairs. These local samples fail to challenge the model to integrate disjointed knowledge, resulting in models proficient at recalling isolated facts but incompetent in complex, multi-hop reasoning.

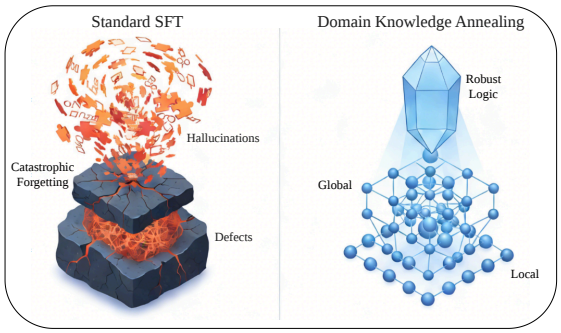


Figure 1: Conceptual comparison of training paradigms. Standard unstructured fine-tuning (Left) causes defects. DKA (Right) adopts an "annealing" schedule, progressing from local facts to global reasoning, to ensure robust knowledge integration.

The lack of a structured curriculum compounds these challenges. Mixing simple factual queries with complex reasoning tasks creates distributional conflicts, causing suboptimal learning dynamics (Figure 1). Such disordered training induces gradient conflicts, hindering the balance between learning atomic facts and high-level reasoning. Moreover, compelling complex reasoning without a solid factual foundation leads to plausible yet erroneous chains (Huang et al., 2024). Additionally, distribution mismatches between synthetic and pre-training data can trigger catastrophic forgetting, compromising general capabilities (Kirkpatrick et al., 2017;

* Corresponding authors.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2665–2682
November 6-10, 2026 ©2026 Association for Computational Linguistics

Zhang et al., 2022; Yang et al., 2024).

To address this, we introduce Domain Knowledge Annealing (DKA)[1], a structured curriculum framework enhancing domain proficiency. Inspired by thermodynamic annealing, DKA organizes training via a "Local-to-Global" progression. It begins with local knowledge injection using intra-document data to absorb high-entropy, discrete concepts. This is followed by a controlled phase guiding the model toward global reasoning, traversing Knowledge Graph topologies to synthesize information across documents. This curriculum mimics annealing, integrating isolated facts into a coherent expert knowledge system.

Methodologically, DKA integrates three core components. First, we construct a fine-grained domain knowledge graph with provenance tracking to map the corpus's logical structure. Leveraging this topology, our data synthesis algorithm utilizes rare node identification and weighted random walks to generate inter-document, multi-hop QA pairs. Finally, we employ a two-stage progressive training regimen aligned with the annealing schedule, minimizing the "alignment tax" by using the base model for data refinement.

DKA also differs from retrieval-based paradigms such as GraphRAG. GraphRAG supplies evidence from an external knowledge base at inference time (Edge et al., 2024), whereas DKA trains the model on provenance-constrained reasoning paths. In our evaluated setting, DKA obtains higher accuracy than our implemented GraphRAG baseline (Section 4.4); this result should not be interpreted as showing that fine-tuning universally dominates retrieval-augmented systems, whose performance depends on retrieval and prompting choices. The two paradigms are complementary: DKA is useful when recurring domain reasoning should be available offline, while GraphRAG can provide updatable external evidence.

Experiments on Linguistics and Law benchmarks demonstrate that DKA surpasses standard random shuffling baselines and achieves the best results among comparably sized models, often outperforming larger general-purpose models. DKA further transfers to open-domain multi-hop question answering (94.66% on the MuSiQue answerable dev split), suggesting that the local-to-global ordering generalizes beyond curated vertical domains. These results reveal that structured data organization, specifically the transition from local facts to global reasoning, is a decisive factor comparable to data scaling for maximizing domain-specific reasoning.

[1]Our code and data are publicly available at https://github.com/zyk42/dka.

2 Related Work

2.1 Data Synthesis

High-quality Supervised Fine-Tuning data is the cornerstone of effective domain adaptation. Yet, manual construction of such data—particularly for multi-document and multi-hop reasoning—remains prohibitively expensive. Early synthesis methods, represented by Self-Instruct (Wang et al., 2023; Jeong, 2024), typically rely on general-purpose priors or intra-document extraction. While scalable, these approaches often yield shallow instruction pairs that fail to challenge a model's ability to synthesize fragmented knowledge (Lu et al., 2025; Susnjak et al., 2025). To address this, recent works have begun to incorporate structural priors. For instance, Synthetic-on-Graph constructs context graphs to capture explicit knowledge associations (Ma et al., 2025), and Condor employs World Knowledge Trees with self-reflection to enhance data quality (Cao et al., 2025). Similarly, WizardLM and Orca focus on progressively increasing instruction complexity or leveraging explanation traces to teach reasoning processes (Xu et al., 2023; Mukherjee et al., 2023).

DKA transcends these approaches by transforming the Knowledge Graph from a static reference into a dynamic engine for reasoning path generation. Unlike Graph of Thoughts, which structures reasoning only during the inference phase (Besta et al., 2024; Yao et al., 2023), DKA internalizes this structure directly into the training data. By integrating rare node identification with weighted random walks, we synthesize complex, multi-hop QA pairs that explicitly require inter-document information synthesis. This fills a critical gap in automated data generation, enabling models to learn "reasoning-by-navigation" rather than simple pattern matching (Song et al., 2025; Yang et al., 2025b).

2.2 Staged Training

Curriculum learning serves as a vital mechanism for adapting LLMs to specialized domains (Xu et al., 2020). While standard fine-tuning often treats data as a uniform mixture, recent findings indicate that data composition and ordering fun-

damentally dictate model capabilities (Dong et al., 2024; Liu et al., 2025b; Wang et al., 2025). Existing solutions typically rely on heuristic curricula, prioritizing tasks based on surface metrics like sequence length (Kim and Lee, 2024) or model perplexity (Luo et al., 2024; Chen et al., 2025). However, these proxies often fail to reflect the true logical density of domain knowledge; a long sequence is not necessarily logically complex, and high perplexity may stem from noise rather than depth. Consequently, curricula based on these shallow metrics often struggle to foster genuine reasoning capabilities, as they do not align with the intrinsic cognitive structure of the domain.

DKA fundamentally diverges from these heuristic approaches by introducing a "Structure-Aware" curriculum rooted in the topological complexity of the domain knowledge graph. Instead of relying on text length or loss signals, DKA defines difficulty through connectivity: progressing from "Local" tasks involving isolated nodes (atomic facts) to "Global" tasks requiring the traversal of multi-hop paths (complex reasoning). This topological progression mirrors the cognitive process of schema acquisition, where learners must internalize discrete concepts before synthesizing them into broader theoretical frameworks. By aligning the training curriculum with the logical structure of knowledge rather than statistical artifacts, DKA ensures a smoother and more robust acquisition of expert-level reasoning.

2.3 Annealing

The metaphor of "annealing" in deep learning has traditionally described learning rate schedules that gradually reduce step sizes to settle models into low-energy minima (Loshchilov and Hutter, 2016). More recently, frontier models have shifted toward "Data Annealing"—dynamically adjusting data distribution during terminal training stages to crystallize capabilities (Dubey et al., 2024; Hu et al., 2024; DeepSeek-AI, 2024). While effective, these methods often treat data annealing as a general quality refinement process, lacking a rigorous structural mapping between the data's logical state and the physical annealing phases. They typically rely on filtering low-quality data rather than actively structuring the thermodynamic evolution of knowledge.

DKA extends this paradigm by defining a rigorous "Knowledge Annealing" process that is strictly governed by domain topology. We posit that the topological transition from isolated entities to interconnected subgraphs provides the most faithful mapping to the thermodynamic cooling process. The first phase corresponds to the injection of high-entropy, local intra-document data, allowing the model to absorb discrete, disordered facts. The second phase then transitions to low-entropy, global inter-document reasoning paths, forcing the model to settle these facts into a stable, ordered logical lattice. This structure-based annealing ensures that the integration of intelligence is not merely a statistical reduction in loss, but a genuine formation of coherent, evidence-grounded reasoning structures.

3 DKA: Domain Knowledge Annealing

This section presents DKA, a generalizable SFT framework designed to construct domain-specific large language models (Figure 2).

The design of DKA is motivated by the need to address inherent limitations in learning from unstructured text. Unstructured corpora often obscure logical relationships between distant concepts; to make these explicit, we introduce a Knowledge Graph that reconstructs such connections. Moreover, since standard training tends to overemphasize frequent concepts, we leverage rare-node-guided synthesis to ensure exposure to the long tail of knowledge. Drawing an analogy to physical annealing—where high thermal energy allows atoms to rearrange before settling into a stable state—we adopt a progressive training strategy that evolves from local knowledge injection to global reasoning integration. DKA integrates three core components: (1) domain knowledge graph construction with provenance tracking, (2) automated synthesis of hierarchical SFT data, and (3) a two-stage "Local-to-Global" progressive training regimen. The overall workflow is outlined in Algorithm 1.

3.1 Domain Knowledge Graph Construction

High-quality, structured knowledge organization is foundational for synthesizing complex training data and mitigating hallucinations (Pan et al., 2024). We begin by processing raw domain texts (e.g., academic papers, professional books) to construct a fine-grained Knowledge Graph. Unlike traditional pipeline-based methods or unified extraction frameworks like UIE (Lu et al., 2022; Yang et al., 2025c), we employ an LLM-driven information extraction approach (Jiang et al., 2024) to identify entities, attributes, and relations within semantically co-

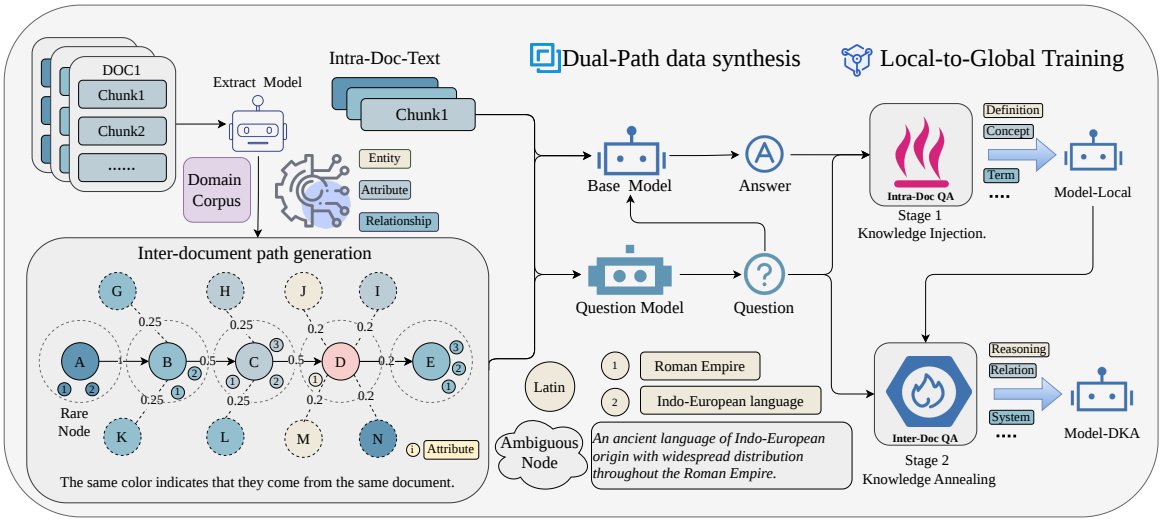


Figure 2: The proposed Domain Knowledge Annealing framework.

Algorithm 1. DKA: Domain Knowledge Annealing.

Input: Domain corpus $\mathcal{D}$, Base LLM M_{base}
Output: Domain-specific LLM M_{DKA}

Phase 1: Knowledge Graph Construction
$\mathcal{C} \leftarrow \text{DocumentChunking}(\mathcal{D})$
$\mathcal{E}, \mathcal{A}, \mathcal{R} \leftarrow \text{InformationExtraction}(\mathcal{C}, M_{\text{extract}})$
$\mathcal{G} \leftarrow \text{BuildGraph}(\mathcal{E}, \mathcal{A}, \mathcal{R})$
for $v \in \mathcal{V}(\mathcal{G})$ **do**
  $v.\text{source} \leftarrow \text{GetSourceChunk}(v)$

Phase 2: Intra-Document Data Synthesis
$\mathcal{Q}_{\text{intra}} \leftarrow \emptyset$
for $c \in \mathcal{C}$ **do**
  $q \leftarrow \text{GenerateQuestion}(M_{\text{question}}, c)$
  $a \leftarrow \text{RefineAnswer}(M_{\text{base}}, q, c)$
  $\mathcal{Q}_{\text{intra}}.\text{append}((q, a))$

Phase 3: Inter-Document Data Synthesis
$\mathcal{V}_{\text{rare}} \leftarrow \text{IdentifyRareNodes}(\mathcal{G})$
$\mathcal{Q}_{\text{inter}} \leftarrow \emptyset$
for $v_{\text{start}} \in \mathcal{V}_{\text{rare}}$ **do**
  $\mathcal{P} \leftarrow \text{WeightedRandomWalk}(\mathcal{G}, v_{\text{start}})$
  for $p \in \mathcal{P}$ **do**
    $p_{\text{amb}} \leftarrow \text{AmbiguifyNode}(p)$
    $q \leftarrow \text{GenerateQuestion}(M_{\text{question}}, p_{\text{amb}})$
    $a \leftarrow \text{RefineAnswer}(M_{\text{base}}, q, p)$
    $\mathcal{Q}_{\text{inter}}.\text{append}((q, a))$

Phase 4: Progressive Training
$M_{\text{Local}} \leftarrow \text{SFT}(M_{\text{base}}, \mathcal{Q}_{\text{intra}})$
$M_{\text{DKA}} \leftarrow \text{SFT}(M_{\text{Local}}, \mathcal{Q}_{\text{inter}})$
return M_{DKA}

herent segments. This results in structured tuples $(\mathcal{E}, \mathcal{A}, \mathcal{R})$, where $\mathcal{E}$ denotes entities, $\mathcal{A}$ represents attributes of entities, and $\mathcal{R}$ stands for relations between entities, with each entity enriched with detailed attribute information (e.g., name, descriptions) to ensure graph integrity.

A defining characteristic of our approach is Provenance Annotation: each node and edge in the graph is explicitly linked to its source text chunk and document ID. This traceability provides the foundational premise for the subsequent inter-document weighted random walk—a critical step for generating diverse, evidence-grounded reasoning trajectories that remain faithful to the original corpus (Sun et al., 2024). Refer to Appendix C for detailed construction protocols.

3.2 Intra-Document QA Pair Generation

Intra-document QA pairs constitute the foundational layer of domain knowledge. For each text chunk, we leverage a more capable, thinking-capable model to formulate a domain-relevant question. To ensure data quality, we apply semantic deduplication to remove redundant questions. Critically, we employ the base model itself to generate the final answers based on the ground-truth chunks. This self-refinement step ensures that the target distribution aligns with the model's inherent linguistic patterns, thereby facilitating efficient learning and mitigating the "alignment tax" typically incurred by distilling from external models.

3.3 Inter-Document QA Pair Generation

To foster inter-document integration and complex reasoning, we design a KG-driven synthesis pipeline that synthesizes multi-hop QA pairs requiring information synthesis from disparate sources. This pipeline consists of two key steps: Rare Node Identification and Inter-Document Path Generation.

3.3.1 Rare Node Identification

To mitigate the bias toward frequent entities, we prioritize "long-tail" concepts. We identify the top $k\%$ of nodes with the highest rarity scores as starting points for reasoning paths, where the parameter k determines the number of inter-document QA pairs to be constructed and is set based on the actual number of nodes in the graph (see Appendix E.2). The rarity score $R(v_i)$ is defined as:

$$R(v_i) = \frac{w(v_i)}{d_{in}(v_i) + d_{out}(v_i)} \tag{1}$$

where $w(v_i)$ represents the semantic weight of the node type, with a default value of 1 for ordinary node types and values greater than 1 for important node types (to amplify their contribution to semantic criticality). d_{in} and d_{out} are in-degree and out-degree, respectively. Notably, $R(v_i)$ is not calculated for isolated nodes—that is, nodes where the in-degree and out-degree are simultaneously zero—and such nodes are directly discarded. This strategy guarantees exposure to structurally sparse but semantically critical concepts.

3.3.2 Inter-Document Path Generation

Starting from a rare node, we generate multi-hop reasoning paths via a weighted random walk. To enforce inter-document reasoning, we introduce a document diversity constraint into the transition probability:

$$P(v_{t+1} \mid v_1, v_2, \ldots, v_t) = \frac{1 + \lambda \cdot \mathcal{I}\{D(v_{t+1}) \notin \mathcal{D}_t\}}{\sum_{v' \in \mathcal{N}(v_t)} [1 + \lambda \cdot \mathcal{I}\{D(v') \notin \mathcal{D}_t\}]} \tag{2}$$

where v_t is the node at step t, $D(v_t)$ denotes the document containing node v_t, $\mathcal{D}_t = \{D(v_1), D(v_2), \ldots, D(v_t)\}$ is the set of documents visited so far, $\mathcal{I}\{\cdot\}$ is the indicator function, $\mathcal{N}(v_t)$ represents the set of nodes connected to v_t, and λ is a hyperparameter that controls both the path length distribution and document diversity. The determination of λ is discussed in the Appendix E.5. Path length is constrained between 3 and 6 hops (where paths with a final length shorter than 3 hops are discarded, and the random walk terminates upon reaching 6 hops) to balance complexity and solvability.

Once a path $P = (v_1, v_2, \ldots, v_k)$ is generated, we collect the attribute information of the nodes in the path (including the extracted text segments for each node). These pieces of information—comprising node attributes and their associated text excerpts—form a multi-document context. Subsequently, we employ the question model to construct a complex question that requires integrating inter-document content from all the collected information in the sequence. Finally, the base model generates the answer based on this path information, ensuring the reasoning process is grounded in the integrated evidence.

3.3.3 Ambiguous Node

To prevent shortcut learning through superficial pattern matching—where models rely on spurious correlations rather than deep understanding (Gururangan et al., 2018)—we introduce ambiguity when constructing questions based on graph paths. Specifically, we mask the explicit entity name and replace it with a set of identifying attributes or a description. For example, instead of asking "What is the relationship between Latin and modern Romance languages?", we might ask "What is the relationship between an ancient language of Indo-European origin with widespread distribution throughout the Roman Empire and modern Romance languages?", provided the description uniquely identifies the node (Latin) within the context. This forces the base model to perform multi-step deduction when answering questions based on the path—first inferring the entity identity from its attributes and then proceeding with the relational reasoning—thereby fostering deeper semantic comprehension.

Additionally, this ambiguity serves as a robustness mechanism, mitigating the impact of potential incompleteness or errors in the Knowledge Graph by forcing the model to rely on semantic context rather than exact graph matching. We further discuss and validate the effectiveness of this strategy through ablation studies in Section 4.5 and a detailed case study in Appendix F.

3.4 Local-to-Global Training as Knowledge Annealing

We implement a "Local-to-Global" two-stage training regimen, which we conceptualize as a Knowledge Annealing process.

Stage 1: Knowledge Injection. The model is fine-tuned on intra-document QA pairs. This phase prioritizes the internalization of domain-specific terminology, definitions, and atomic facts, estab-

lishing a robust knowledge base. By focusing on local context first, the model absorbs vast amounts of discrete, disordered atomic facts (High Entropy regarding the structure of knowledge, i.e., a bag of facts), analogous to the phase where atoms gain energy and mobility.

Stage 2: Knowledge Annealing. The model is further fine-tuned on inter-document, multi-hop QA pairs. Building upon the internalized knowledge, this stage acts as the annealing phase, forcing these facts to establish strong logical connections and reducing logical uncertainty (Low Entropy regarding the reasoning path, i.e., a structured lattice). This high-quality, complex data anneals the model, integrating the fragmented knowledge into a structured, expert-level reasoning capability.

This structured, progressive curriculum mirrors the principles of thermal annealing: by first elevating and then carefully tempering the model's "cognitive state," we avoid the instability that arises from learning atomic facts and complex relations concurrently. The result is the systematic organization of initially fragmented knowledge into a stable, crystalline reasoning faculty. Note that we use "annealing" as a name for this structure-defined curriculum rather than as a new optimizer or an explicit temperature schedule: the stage transition is determined by provenance and graph connectivity, not by a learning-rate schedule, sequence length, or perplexity. The staged ordering itself is the testable design choice, and Section 4.2 reports diagnostic measurements supporting it.

4 Experiments

We evaluate the DKA framework across two distinct domains: Linguistics and Law. In addition, we examine whether the approach transfers to open-domain multi-hop question answering, using the Wikipedia paragraphs of the MuSiQue training split (Trivedi et al., 2022) as the corpus and its answerable dev questions for evaluation (corpus details in Appendix A.3). To isolate the impact of our data composition and training strategy, we enforce strict consistency in hyperparameters and hardware environments across all experiments (detailed in Appendix E). To mitigate the impact of random noise, all reported results are the average of 5 independent evaluation runs.

We first analyze the complexity of our synthesized data. Figure 3 illustrates the token length distributions for the Linguistics domain. Inter-document samples exhibit significantly longer sequences in both questions and answers compared to intra-document samples, reflecting the increased reasoning depth required to synthesize information from disparate sources.

Strategy	Intra-Doc	Inter-Doc	Local-to-Global
Local FT	✓	×	×
Mix FT	✓	✓	×
DKA FT	✓	✓	✓

Table 1: Comparison of different training strategies.

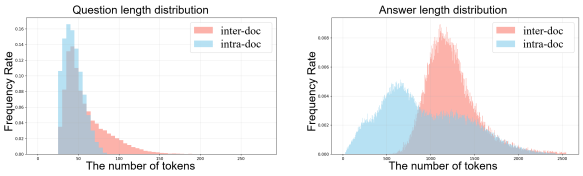


Figure 3: Length distributions of intra-document/inter-document questions and answers.

To benchmark DKA, we compare three distinct training strategies outlined in Table 1: (1) Local FT, trained exclusively on intra-document data; (2) Mix FT, trained on a shuffled mixture of intra- and inter-document data; and (3) DKA FT, our proposed progressive approach that evolves from local knowledge injection to global reasoning integration.

4.1 Effectiveness of DKA

4.1.1 Linguistics Domain Analysis

We first utilized the DKA framework to train a linguistics-specific model based on Qwen3-8B. We evaluated its performance on Lingua-Bench, a comprehensive benchmark we developed. It covers diverse linguistics tasks (Xu et al., 2025), categorized by difficulty into "Easy" (factual recall) and "Challenging" (multi-hop reasoning). We compared it with strong baselines including GLM4-9B/32B (GLM et al., 2024), Qwen3 series (Yang et al., 2025a), GPT-4o (Achiam et al., 2023), and DeepSeek-V3.2Exp (Liu et al., 2025a).

Results in Table 2 demonstrate the effectiveness of our method. It can be observed that Qwen3-8B (63.15%) outperforms GLM4-9B (62.64%), justifying our choice of Qwen3-8B as the base model, with model performance generally improving as parameter size increases. Our Qwen3-8B + DKA FT model, with a score of 70.87%, ranks second only to DeepSeek-V3.2Exp and surpasses the closed-source GPT-4o (69.37%). Notably, in challenging tasks, Qwen3-8B + DKA FT (True/False 62.73%, Single Choice 47.54%) outperforms DeepSeek-V3.2Exp (True/False 62.36%,

Model	Easy		Challenging		Avg.
	True/False	Single Choice	True/False	Single Choice	
GLM4-9B	66.06	68.60	55.13	39.34	62.64
Qwen3-30B-A3B	63.13	77.71	57.03	40.98	65.15
Qwen3-32B	69.83	77.71	58.56	42.62	68.20
GLM4-32B	67.32	76.60	57.03	37.70	66.09
GPT-4o	71.23	78.18	60.83	43.72	69.37
DeepSeek-V3.2Exp	**74.02**	**80.38**	62.36	46.45	**71.76**
Qwen3-8B	67.18	71.66	50.19	36.43	63.15
+ Local FT	72.62	78.02	57.03	40.43	68.99
+ Mix FT	72.91	77.71	61.22	45.36	70.10
+ DKA FT	**73.04**	**78.49**	**62.73**	**47.54**	**70.87**

Note: Avg represents the ratio of all correct answers to the total number of questions

Table 2: Performance comparison on Lingua-Bench (Accuracy %).

Single Choice 46.45%). The superior performance of Qwen3-8B + DKA FT, especially on challenging tasks, confirms that DKA successfully bridges the gap between general model capabilities and domain-specific reasoning.

4.1.2 Legal Domain Analysis

To validate the generalizability of DKA, we trained legal domain models based on Qwen2.5-7B and Qwen3-8B, and tested them on the DISC-Law-Eval benchmark (Yue et al., 2023). First, we applied DKA to Qwen2.5-7B, a model with a strong initial capability in the legal domain (61.45%). Results (Table 3) show that DKA achieved a substantial gain, reaching 67.62%. This performance ranks second only to the massive DeepSeek-V3.2Exp (70.11%) and surpasses larger models like GLM4-32B (63.13%). To further verify the robustness of DKA and determine if it is effective only on strong base models, we applied our framework to Qwen3-8B, which has a relatively weaker baseline (53.88%). Remarkably, applying DKA yielded a dramatic improvement, boosting the score to 63.71%—an increase of nearly 10 percentage points. These experiments reveal a critical insight: across both the stronger and the weaker base model, DKA consistently provides significant performance uplifts.

4.1.3 Ablation on DKA

To further verify the efficacy of DKA, we tested the models trained using the Local FT and Mix FT methods. In the linguistic domain, Table 2 shows that Qwen3-8B + Local FT, which was trained on intra-document data, achieved an accuracy of 68.99%, surpassing Qwen3-32B (68.20%). Its performance on "Easy" tasks and "Challenging" True/False questions improved significantly (by more than 5%), confirming the necessity of domain-specific fine-tuning. Qwen3-8B + Mix FT, which incorporates inter-document data during training, further improved performance on "Challenging" tasks (an approximate 5% increase) but showed stagnation on "Easy" tasks; notably, its performance on single-choice questions (77.71%) even lagged behind Qwen3-8B + Local FT (78.02%). This indicates that conflicting learning objectives can impede model performance. Qwen3-8B + DKA FT (70.87%), trained with the DKA framework, outperformed both Qwen3-8B + Local FT (68.99%) and Qwen3-8B + Mix FT (70.10%) across all evaluation metrics, which corroborates the effectiveness of sequential learning.

Similarly, in the legal domain, DKA outperformed both the Local and Mix strategies regardless of whether it was applied to a stronger or weaker base model. For example, Qwen2.5-7B + DKA FT (67.62%) outperformed Qwen2.5-7B + Mix FT (66.45%), and Qwen3-8B + DKA FT (63.71%) outperformed Qwen3-8B + Mix FT (62.43%). These results are consistent with the benefit of the structured ordering in our evaluated settings, but do not establish that the framework is model-agnostic or universally optimal.

4.2 Training Dynamics Analysis

To investigate training dynamics, we compared the loss curves of DKA FT and Mix FT strategies. Figure 4 shows the training loss for the linguistics domain model (loss for the legal domain model is in the Appendix E.4). While both strategies show similar initial declines, the DKA FT approach initiates a new, deeper convergence phase in the second stage, ultimately reaching a lower stable loss.

Model	Single Choice						Multiple Choice				Avg.
	NJE	PAE	CPA	UNGEE	PFE	LBK	NJE	PAE	CPA	UNGEE	
LexiLaw	20.11	23.73	24.87	31.56	31.76	40.36	7.56	10.14	19.17	16.09	21.50
LawGPT	22.91	31.36	25.38	30.31	34.71	29.09	6.26	7.61	16.67	13.79	20.60
ChatLaw	27.56	31.36	35.53	35.62	42.35	41.09	7.99	9.42	11.67	17.24	25.20
DISC-LawLLM	42.09	40.68	39.59	50.94	57.06	54.91	19.87	18.48	19.17	25.29	37.10
GLM4-9B	53.82	55.93	69.54	74.38	80.00	73.09	27.00	22.10	50.83	57.47	53.22
Qwen3-30B-A3B	56.61	61.02	80.71	79.69	85.88	82.91	34.56	35.14	59.17	66.67	60.48
GLM4-32B	61.82	57.63	80.20	75.86	87.65	81.45	35.42	38.41	70.83	**75.86**	63.13
DeepSeek-V3.2Exp	**67.23**	**68.64**	**86.80**	**91.25**	92.35	**92.36**	**41.90**	48.91	**73.33**	73.56	**70.11**
Qwen3-8B	54.75	56.78	72.59	77.81	78.82	76.00	24.75	25.72	43.33	55.17	53.88
+ Local FT	55.68	56.78	74.11	81.56	82.94	79.27	30.67	36.59	51.56	62.07	58.17
+ Mix FT	58.29	58.47	77.16	82.19	**85.88**	82.18	38.66	43.84	56.67	65.52	62.43
+ DKA FT	**59.22**	**58.47**	**77.66**	**84.69**	84.71	**83.27**	**39.96**	**46.74**	**63.33**	**67.82**	**63.71**
Qwen2.5-7B	58.85	55.08	81.73	80.62	90.59	81.09	33.48	39.86	68.33	58.62	61.45
+ Local FT	61.64	71.19	79.19	81.88	83.53	82.18	38.88	47.46	65.83	66.67	64.03
+ Mix FT	63.69	68.64	82.23	83.13	91.18	84.36	39.96	50.00	66.67	65.52	66.45
+ DKA FT	**64.80**	**68.64**	**84.26**	**84.06**	**92.94**	**86.91**	**40.17**	**51.45**	**70.83**	**67.82**	**67.62**

Note: Avg represents the ratio of all correct answers to the total number of questions.
NJE: National Judicial Exam; PAE: Patent Agent Exam; CPA: Certified Public Accountant; LBK: Legal Basics Knowledge; UNGEE: Unified National Graduate Entrance Examination; PFE: Professional Front-end Examination.

Table 3: Performance comparison on DISC-Law-Eval benchmark (Accuracy %).



Figure 4: Comparison of training loss curves based on Qwen3-8B (Linguistics). The DKA FT strategy converges to a lower loss than Mix FT strategy.

Model	MuSiQue Acc. (%)
Qwen2.5-7B	85.30
+ Local FT	88.43
+ Mix FT	91.27
+ DKA FT	**94.66**

Table 4: Transfer evaluation on the answerable dev split of MuSiQue.

Beyond the two vertical domains above, we also examine whether the local-to-global ordering transfers to a noisier open-domain setting: multi-hop question answering over Wikipedia paragraphs from MuSiQue (Trivedi et al., 2022) (benchmark details in Appendix B.4). As shown in Table 4, with Qwen2.5-7B, DKA FT (94.66%) outperforms Mix FT (91.27%) by 3.39 points on the answerable dev questions, as scored by the LLM judge, suggesting that the local-to-global ordering transfers beyond vertical domains.

To probe the mechanism behind this second convergence phase, we recorded two diagnostics on the MuSiQue corpus (Table 5). First, the average token entropy decreases from 1.4108 in DKA Stage 1 to 1.3062 in Stage 2, whereas Mix FT remains at 1.4452, consistent with the intended transition from local factual supervision to structured relational supervision. Second, we measured the cosine similarity between gradient vectors of batch pairs sampled every 50 steps: 0.6736 for DKA Stage 1 (intra–intra), 0.1580 for DKA Stage 2 (inter–inter, after Stage 1), and -0.0523 for Mix FT (intra–inter). The negative similarity under Mix FT indicates that jointly optimizing the two supervision types produces misaligned gradients, which DKA avoids by separating them into consecutive stages. We present these measurements as diagnostic evidence rather than complete causal proof; a within-type control under Mix FT is left to future work.

Setting	Avg. entropy	Grad. cos sim
DKA Stage 1	1.4108	0.6736
DKA Stage 2	1.3062	0.1580
Mix FT	1.4452	-0.0523

Table 5: Diagnostic training statistics on the MuSiQue corpus: average token entropy and mean cosine similarity between gradients of sampled batches.

4.3 Impact on General Capabilities

A natural concern with domain-specific fine-tuning is catastrophic forgetting. To examine this, we evaluated the trained models on MMLU (Hendrycks et al., 2020), SuperGPQA (Team, 2025), and BBH (Suzgun et al., 2023) (Table 6). All DKA vari-

Model	MMLU	SuperGPQA	BBH
Qwen3-8B	76.89	31.64	78.40
+ Local FT (Lingua)	76.02	30.14	76.21
+ Mix FT (Lingua)	73.21	28.06	74.88
+ DKA FT (Lingua)	75.55	29.88	75.56
+ Local FT (Law)	75.86	29.87	75.94
+ Mix FT (Law)	72.58	27.92	74.63
+ DKA FT (Law)	74.54	29.74	75.93
Qwen2.5-7B	74.16	26.34	70.04
+ Local FT (Law)	71.06	25.96	67.82
+ Mix FT (Law)	70.02	24.93	65.25
+ DKA FT (Law)	70.65	25.96	66.95
+ Local FT (MuSiQue)	72.23	26.34	68.37
+ Mix FT (MuSiQue)	70.34	25.12	65.83
+ DKA FT (MuSiQue)	71.82	26.14	67.86

Table 6: Performance on general benchmarks.

Domain	Model	GraphRAG	DKA FT
Linguistics	Qwen3-8B	66.04	**70.87**
Law	Qwen3-8B	56.07	**63.71**
Law	Qwen2.5-7B	63.17	**67.62**
MuSiQue	Qwen2.5-7B	89.83	**94.66**

Table 7: GraphRAG vs. DKA (Acc %).

ants remain close to their base models, indicating that DKA does not induce catastrophic forgetting. Moreover, DKA consistently preserves general capabilities better than Mix FT: on Linguistics, DKA scores 75.55/29.88/75.56 versus 73.21/28.06/74.88 for Mix FT, and on Law 74.54/29.74/75.93 versus 72.58/27.92/74.63. We attribute this advantage to the staged schedule, which absorbs domain knowledge progressively rather than exposing the model to an abrupt heterogeneous mixture.

4.4 Comparison with GraphRAG

To verify KG quality and the necessity of fine-tuning, we compare the base model with its GraphRAG-equipped variant. Table 7 shows that GraphRAG brings consistent gains over the base model (Tables 2, 3, and 4) across all domains, confirming that our KG provides high-quality domain knowledge missing from the base model. More importantly, DKA FT further surpasses GraphRAG by a clear margin in every setting, though conditional on the particular graph and retriever.

4.5 Ablation on Ambiguous Node

To validate the ambiguous node strategy (Section 3.3.3), we compare standard DKA against a variant whose second stage uses explicit entity names in cross-document QA pairs. Table 8 shows that ambiguous nodes consistently outperform the explicit variant in all settings; without the ambiguity challenge, the model may resort to superficial entity-name matching rather than "reasoning-by-navigation."

Domain	Model	Explicit	Ambiguous
Linguistics	Qwen3-8B	69.53	**70.87**
Law	Qwen3-8B	63.23	**63.71**
Law	Qwen2.5-7B	66.94	**67.62**
MuSiQue	Qwen2.5-7B	94.18	**94.66**

Table 8: Ablation on the ambiguous node strategy (Acc %).

5 Conclusion

We introduce Domain Knowledge Annealing (DKA), a paradigm that transforms domain adaptation from simple data exposure into rigorous knowledge integration. By integrating a provenance-aware knowledge graph with a "Local-to-Global" curriculum, DKA guides models from absorbing atomic facts to synthesizing stable reasoning structures, mirroring thermodynamic annealing. Evaluations across linguistics and law, including the novel Lingua-Bench and a transfer study on open-domain multi-hop QA (MuSiQue), show that DKA achieves state-of-the-art results among comparably sized models, often outperforming much larger general-purpose models. These findings reveal that the structural organization of training data is as decisive as its quantity, paving the way for efficiently creating trustworthy expert LLMs in high-stakes domains.

Acknowledgments

This study is supported by the National Social Science Foundation of China (No. 24&ZD186).

Limitations

DKA has several limitations. First, synthetic supervision inherits noise, omissions, and biases from the source corpus and automatic graph extraction; our node-removal study shows moderate robustness to missing nodes, but relation-level errors and adversarial graph construction remain untested. Second, temporally inconsistent or contradictory documents may be merged into the same graph, yielding conflicting paths; corpus versioning and conflict-aware construction mitigate but do not solve this, and dynamic maintenance remains open. Finally, DKA adds graph-construction and QA-synthesis cost and may not suit settings where evidence changes frequently. These limitations motivate future work on incremental updates, conflict resolution, and evaluation on noisier domains.

Ethical Considerations

Developing domain-specific large language models carries ethical responsibilities. Our models, particularly legal ones, are intended for research and assistive purposes only, not as a substitute for professional legal advice. Bias inherited from the source corpora remains a concern: we filter harmful content during data synthesis, but cannot guarantee its complete absence. Generated reasoning chains could also be misused to produce plausible-sounding misinformation, so deployment should include human oversight and safety guardrails.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17682–17690. AAAI Press.

Maosong Cao, Taolin Zhang, Mo Li, Chuyu Zhang, Yunxin Liu, Haodong Duan, Songyang Zhang, and Kai Chen. 2025. Condor: Enhance LLM alignment with knowledge-driven data synthesis and refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22392–22412, Vienna, Austria. Association for Computational Linguistics.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. 2025. Self-evolving curriculum for llm reasoning. *arXiv preprint arXiv:2505.14970*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Y. Ge, W. Hua, K. Mei, and 1 others. 2023. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, and 1 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

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

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, and 1 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1428, Bangkok, Thailand. Association for Computational Linguistics.

Cheonsu Jeong. 2024. Fine-tuning and utilization methods of domain-specific llms. *arXiv preprint arXiv:2401.02981*.

Shu Jiang, Zuchao Li, Hai Zhao, and Weiping Ding. 2024. Entity-relation extraction as full shallow semantic dependency parsing. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:1088–1099.

Jisu Kim and Juhwan Lee. 2024. Strategic data ordering: Enhancing large language model performance through curriculum learning. *arXiv preprint arXiv:2405.07490*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. Knowledge boundary of large language models: A survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157, Vienna, Austria. Association for Computational Linguistics.

Yucheng Li, Shuai Wang, Hongzhou Ding, and 1 others. 2023. Large language models in finance: A survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 374–382.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Kai Liu, Ze Chen, Zhihang Fu, Wei Zhang, Rongxin Jiang, Fan Zhou, Yaowu Chen, Yue Wu, and Jieping Ye. 2025b. Structure-aware domain knowledge injection for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29443–29464, Vienna, Austria. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Rui Lu, Zhenyu Hou, Zihan Wang, and 1 others. 2025. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl. *arXiv preprint*, arXiv:2509.10446.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Yunfei Luo, Yuyang Liu, Rukai Cai, and Tauhidur Rahman. 2024. Start simple: Progressive difficulty multitask learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 48–55, Mexico City, Mexico. Association for Computational Linguistics.

Shengjie Ma, Xuhui Jiang, Chengjin Xu, and 1 others. 2025. Synthesize-on-graph: Knowledgeable synthetic data generation for continue pretraining of large language models. *arXiv preprint*, arXiv:2505.00979.

Shervin Minaee, Tomas Mikolov, Nal Nikzad, and 1 others. 2024. Large language models: A survey. *arXiv preprint*, arXiv:2402.06196.

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

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. 2025. Injecting domain-specific knowledge into large language models: A comprehensive survey. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25297–25311, Suzhou, China. Association for Computational Linguistics.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Zihan Sun. 2023. A short survey of viewing large language models in legal aspect. *arXiv preprint*, arXiv:2303.09136.

T. Susnjak, P. Hwang, N. Reyes, and 1 others. 2025. Automating research synthesis with domain-specific large language model fine-tuning. *ACM Transactions on Knowledge Discovery from Data*, 19(3):1–39.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

M-A-P Team. 2025. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. *CoRR*, abs/2502.14739.

Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kabilan Elangovan, and 1 others. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. ♫ MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Bin Wang, Chaochao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. Mineru: An open-source solution for precise document content extraction. *ArXiv*, abs/2409.18839.

Suqing Wang, Zuchao Li, Luohe Shi, Bo Du, Hai Zhao, Yun Li, and Qianren Wang. 2025. From parameters to performance: A data-driven study on LLM structure and development. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 26084–26101. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

LiQing Xu, Qiwei Li, Tianshuo Peng, Zuchao Li, Hai Zhao, and Ping Wang. 2025. Can large language models be good language teachers? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 23957–23971. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Shengjie Ma, Aofan Liu, Hui Xiong, and Jian Guo. 2025b. LongFaith: Enhancing long-context reasoning in LLMs with faithful synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3236–3256, Vienna, Austria. Association for Computational Linguistics.

Lu Yang, Jiajia Li, En Ci, Lefei Zhang, Zuchao Li, and Ping Wang. 2025c. Label drop for multi-aspect relation modeling in universal information extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers*, pages 5021–5040. Association for Computational Linguistics.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *CoRR*, abs/2305.16582.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

T. Zhang, X. Wang, B. Liang, and 1 others. 2022. Catastrophic interference in reinforcement learning: A solution based on context division and knowledge distillation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9925–9939.

Wayne Xin Zhao, Kun Zhou, Junyi Li, and 1 others. 2023. A survey of large language models. *arXiv preprint*, arXiv:2303.18223.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17744–17761, Suzhou, China. Association for Computational Linguistics.

Appendix

A Corpus Details

A.1 Linguistics Corpus

We collected 22,966 linguistics papers and 150 related professional books (all publicly available and covering both Chinese and English), spanning various branches of linguistics such as syntax, phonology, morphology, semantics, psycholinguistics, sociolinguistics, and cognitive linguistics. For texts that could not be directly converted to Markdown format, we used the MinerU tool for Optical Character Recognition (OCR) processing (Wang et al., 2024). We cleaned the text data by removing irrelevant content such as publisher information, page numbers, and references, as well as personal privacy-related information.

A.2 Chinese Legal Corpus

The legal corpus is derived from open-source data, which exclusively consists of Chinese-language data and includes current Chinese legal texts as well as public corpora in the legal domain, with all personally sensitive information removed during the preprocessing stage.

A.3 MuSiQue Corpus

For the open-domain transfer setting, we use the Wikipedia paragraphs of the MuSiQue training split (Trivedi et al., 2022) as the corpus (71,090 paragraphs). Following the same construction pipeline as in the main domains, we build a knowledge graph of 100,758 nodes and 105,140 edges over this corpus and synthesize 17,623 intra-document and 9,040 inter-document QA pairs. Evaluation is conducted on the 2,417 answerable questions of the MuSiQue dev split with an LLM-as-a-judge protocol (Appendix B.4).

B Benchmark Statistics

B.1 Lingua-Bench Statistics

Table 9 summarizes the composition of Lingua-Bench, the first comprehensive benchmark specifically designed for the linguistics domain. Unlike existing benchmarks that focus on general language understanding, Lingua-Bench rigorously evaluates domain-specific expertise across core subfields including syntax, phonology, morphology, semantics, psycholinguistics, sociolinguistics, and cognitive linguistics.

The dataset comprises 1,799 questions, categorized into two types (True/False and Single Choice) and two difficulty levels. "Easy" questions (1,353 items) are curated from graduate entrance examinations, assessing fundamental factual recall and concept definition. "Challenging" questions (446 items) are synthesized via an inter-document knowledge graph to enforce multi-hop reasoning, requiring the model to synthesize information from disjoint sources rather than relying on rote memorization. This stratified design fills a critical gap in evaluating expert-level linguistic reasoning.

Type	Difficulty	Count
True/False	Easy	716
True/False	Challenging	263
Single Choice	Easy	637
Single Choice	Challenging	183
Total		**1799**

Table 9: Statistics of Lingua-Bench.

B.2 Legal Benchmark Statistics

The Legal Benchmark utilizes the DISC-Law-Eval dataset, a comprehensive evaluation suite designed to assess legal reasoning capabilities across diverse difficulty levels and subject matters. As detailed in Table 10, the benchmark encompasses a wide spectrum of legal expertise, ranging from fundamental concepts (Legal Basics Knowledge, LBK) to highly specialized professional qualifications: the National Judicial Exam (NJE), Patent Agent Exam (PAE), Certified Public Accountant (CPA) exam, and the legal track of the Unified National Graduate Entrance Examination (UNGEE), plus the benchmark's internal Professional Front-end Examination (PFE) subset. This multi-layered structure allows for a fine-grained analysis of the model's proficiency in handling both routine legal queries and complex, high-stakes case analysis.

Subject	Difficulty	Single Choice	Multiple Choice	Total
NJE	Hard	537	463	1000
PAE	Hard	118	276	394
CPA	Hard	197	120	317
UNGEE	Medium	320	87	407
LBK	Easy	275	-	275
PFE	Easy	170	-	170
Total		**1617**	**946**	**2563**

Table 10: Statistics of Legal Benchmark (DISC-Law-Eval).

B.3 General Benchmarks

We evaluate general capabilities using the following standard benchmarks:

- **MMLU (Massive Multitask Language Understanding)**: A comprehensive test of general knowledge and problem-solving abilities across 57 subjects.
- **SuperGPQA (Graduate-Level Google-Proof Q&A)**: A challenging dataset designed to test graduate-level reasoning and domain-specific knowledge.
- **BBH (Big-Bench Hard)**: A subset of Big-Bench tasks that focuses on complex reasoning capabilities where language models typically struggle.

These benchmarks serve as a control group to monitor catastrophic forgetting, ensuring that our domain-specific optimization does not compromise the model's fundamental reasoning and general knowledge base.

B.4 MuSiQue Benchmark

MuSiQue (Trivedi et al., 2022) is an open-domain multi-hop QA benchmark built over Wikipedia, in which each question is composed bottom-up from interconnected single-hop questions and is explicitly filtered so that it cannot be solved by single-hop shortcuts. Questions span 2–4 reasoning hops and cover diverse reasoning structures (e.g., sequential chains and branching compositions), making it a stringent test of cross-document reasoning. We use the answerable version (MuSiQue-Ans). Since the official test annotations are not publicly released, we evaluate on the 2,417 answerable questions of the dev split. Rather than the official EM/F1 metrics, we adopt an LLM-as-a-judge protocol: a strong LLM compares each generated answer against the gold answer and judges its correctness, which better accommodates free-form, generative responses.

C Knowledge Graph Extraction

The construction of our domain knowledge graph relies on LangExtract, an open-source structured text extraction library developed by Google. This library supports the extraction of structured text by configuring specific extraction examples, and we have defined two extraction types: entities and relations.

For the entity type, the extracted information includes the entity name, the source document where the entity is located, and the relevant attributes of the entity. For the relation type, the extracted content covers the relation name, the name of the head entity involved in the relation, the name of the tail entity, and the corresponding source document of the relation.

The primary reason for adopting LangExtract lies in its document traceability capability: it can label the source document of each extracted entity and relation, which lays a solid foundation for inter-document weighted random walk. The extracted structured data can be converted with minimal effort and then imported into Neo4j, on which our knowledge graph is built. Subsequent tasks, including rare node identification and inter-document weighted random walk, are also implemented by leveraging the GraphDatabase module of the Python Neo4j library.

D Prompt Templates

D.1 Single-Document Data Synthesis

Question Generation Prompt

Description: You are an expert in [Domain] text analysis and question design, capable of extracting key [Domain] information from complex paragraphs and generating high-quality questions to fine-tune datasets.
Reference Content:
{text}
Constraints:
1.Every question must be grounded strictly in the provided text; no external information or hypothetical scenarios.
2.Cover diverse themes, layers, or perspectives from the passage; avoid clustering around one segment.
3.Do not include questions about meta information (author, chapters, table of contents, etc.).
4.Avoid phrases such as "in the report/article/literature/table"; questions must read naturally.

Answer Generation Prompt

Description: You are an expert in generating fine-tuning datasets, skilled at generating accurate answers to questions from the given content, ensuring the accuracy and relevance of the answers.
Reference Content:
{text}
Question:
{question}
Constraints:
1.The answer must be based on the given content.
2.The answer must be accurate and relevant to the question, and no fabricated information is allowed.
3.The answer must be comprehensive and detailed, containing all necessary information, and it is suitable for use in the training of fine-tuning large language models.

This section details the prompt templates em-

ployed in our DKA framework for automated data synthesis. These prompts are designed to guide the base model in generating high-quality, domain-specific Q&A pairs. We categorize them into two distinct phases: Intra-Document Data Synthesis, which focuses on extracting atomic knowledge from individual texts, and Inter-Document Data Synthesis, which leverages the Knowledge Graph to construct complex, multi-hop reasoning tasks.

D.2 Inter-Document Data Synthesis

Question Generation Prompt

You are a senior expert in [Domain], tasked with designing a challenging multi-hop reasoning question for an advanced [Domain] course. Based on the following knowledge graph information, craft a question centered around the "rare_node" that requires at least three reasoning steps to answer.
Knowledge graph information:
{record_json}
Design requirements:
1. Obscure key details (e.g., names of scholars, dates, theory names, specific research objects, etc.) to increase difficulty.
2. Do not explicitly mention all critical elements from the knowledge chain; instead, hint at them indirectly.
3. The question must require at least three distinct reasoning steps to arrive at the correct answer.
4. The question should be grounded in professional [Domain] context, free of factual errors, and must have a single unambiguous answer.
5. The question must focus on the "rare_node," and the correct answer must directly relate to it.
6. Do not include phrases such as "according to the provided information," "based on the knowledge graph," or "given the reference data."
Output only the question itself—no explanations, no answers, and no additional text.

Answer Generation Prompt

You are a senior expert in [Domain], tasked with answering a highly challenging multi-hop reasoning question derived from a [Domain] knowledge graph.
Knowledge graph path information: {record_info}
Question: {question}
Instructions for your response:
1. The answer is embedded within the provided knowledge graph path, but you must articulate a clear, step-by-step reasoning process to arrive at it.
2. Your reasoning should be thorough, logically sound, and grounded in [Domain] principles.
3. Do not mention phrases like "according to the knowledge graph," "based on the graph path," or any reference to external graph data.
4. You already possess this information as part of your expertise—frame all reasoning as coming from your own knowledge (e.g., "Based on my knowledge," "From what I understand," "Recalling my knowledge").
5. Avoid any wording that suggests you are consulting, retrieving, or referencing an external graph. Instead, present the information as inherent to your expert understanding of [Domain].

E Experiment

E.1 Experimental Settings

We utilized LLaMA-Factory as our experimental framework, employing the same training configuration detailed below for both the linguistics and legal domain models. The specific configuration is as follows:

- **Hardware**: 4 × 32GB GPUs
- **Batch Size**: 1 per GPU (Data Parallel)
- **Learning Rate**: 5×10^{-5}
- **LoRA Alpha**: 16
- **LoRA Rank**: 8
- **LR Scheduler**: Cosine
- **Max Grad Norm**: 1.0
- **Num Train Epochs**: 2.0
- **Optimizer**: AdamW

E.2 SFT Data Construction Details

To construct a high-quality SFT dataset, we employed a stratified sampling strategy based on the domain Knowledge Graph (KG), ensuring coverage of both fundamental facts and complex reasoning paths. For the Linguistics domain (410,533 nodes), we first sampled 100,000 intra-document QA pairs to establish a broad knowledge base. To enhance reasoning depth, we identified "rare nodes" (top 15% rarity) as anchors for weighted random walks, from which 44,006 valid inter-document paths were generated (not every anchor yields a walk satisfying the 3–6 hop constraint). Similarly, for the Legal domain (180,918 nodes), we combined 60,000 intra-document QAs with 23,031 valid inter-document paths derived from top 20% rare anchors. This dual-layered approach guarantees that the model masters both atomic knowledge and multi-hop reasoning capabilities.

Construction Efficiency. DKA achieves high efficiency through parallel processing. Using a locally deployed Qwen2.5-7B via vLLM (16 concurrency) as the extraction model, we attained an extraction rate of approximately 15,000 nodes per hour. Consequently, constructing the Linguistics graph took about 30 hours, while the Legal and MuSiQue graphs required 14 and 7 hours, respectively (Table 11). Notably, graph construction is

significantly faster than QA synthesis, particularly compared to reasoning-intensive models. Since the subsequent random walk process is computationally negligible on CPU, DKA introduces only marginal overhead to the overall data preparation pipeline. The final dataset statistics are summarized in Table 12.

Domain	Total Nodes	Time (h)	Extraction Model
Linguistics	410,533	30	Qwen2.5-7B
Legal	180,918	14	Qwen2.5-7B
MuSiQue	100,758	7	Qwen2.5-7B

Table 11: Statistics of Knowledge Graph Construction.

Domain	Intra-Document QA	Inter-Document QA	Total
Linguistics	100,000	44,006	144,006
Legal	60,000	23,031	83,031
MuSiQue	17,623	9,040	26,663

Table 12: Scale of the Supervised Fine-Tuning (SFT) Datasets.

E.3 QA Token Distribution Analysis

We analyzed the token length distributions of the synthesized QA pairs across all three domains (Linguistics, Legal, and MuSiQue). The question length distributions show a diverse range of query complexities, while the answer length distributions confirm that our models are trained to generate comprehensive, detailed responses rather than short, extractive answers. This pattern holds consistently across the three domains and is crucial for training models to perform substantial reasoning and explanation.

E.4 Training Dynamics and Convergence Analysis

We analyzed the training dynamics of DKA FT and standard Mix FT. Figures 5 and 6 show the training loss curves for the legal domain and the MuSiQue corpus, respectively (both with Qwen2.5-7B). A key observation is the distinct behavior of the DKA strategy (bottom curve). While both DKA FT and the baseline (Mix FT) exhibit similar initial loss reduction, DKA FT initiates a second, deeper convergence phase during the transition from the "local" (intra-document) stage to the "global" (inter-document) stage. This "annealing" effect—where the model first solidifies basic knowledge before refining complex connections—allows DKA FT to escape local minima and ultimately reach a lower stable loss. This confirms that our structured, progressive curriculum facilitates more effective optimization than random data mixing.

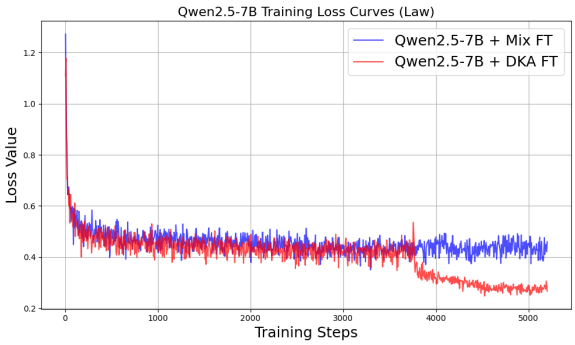


Figure 5: Comparison of training loss curves based on Qwen2.5-7B (Law). The DKA FT strategy converges to a lower loss than Mix FT.

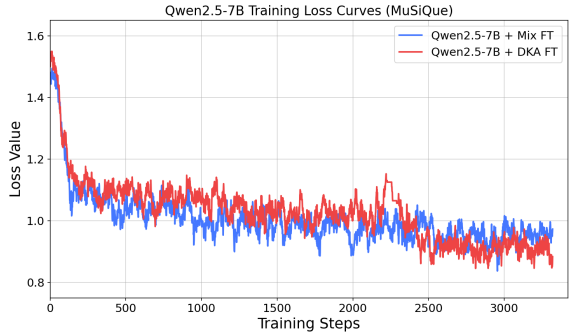


Figure 6: Comparison of training loss curves based on Qwen2.5-7B (MuSiQue). The DKA FT strategy converges to a lower loss than Mix FT.

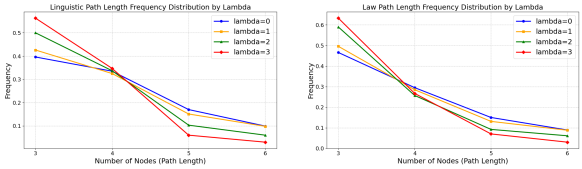


Figure 7: The effect of different λ values on link length distribution for Linguistics (left) and Legal (right) domains.

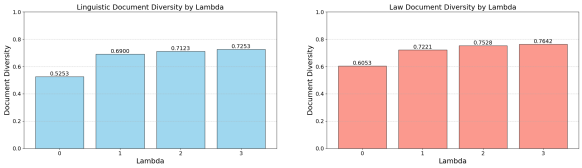


Figure 8: The effect of different λ values on document diversity for Linguistics (left) and Legal (right) domains.

E.5 Hyperparameter Determination

The parameter λ weights the random walk toward unseen documents in our weighted random walk algorithm, balancing exploration of new documents against staying within the current context. Its value critically influences the structural properties of generated reasoning paths, affecting both path length distribution and document diversity.

We define **Document Diversity** to quantify the

information breadth of a reasoning path:

$$Diversity = \frac{1}{N}\sum_{i=1}^{N}\frac{|\{\mathcal{D}(v) \mid v \in P_i\}|}{|P_i|} \quad (3)$$

where N denotes the total reasoning paths, $|P_i|$ is path length, and $|\{\mathcal{D}(v) \mid v \in P_i\}|$ is the count of distinct documents in path P_i. Higher diversity indicates the chain aggregates information from wider sources, essential for synthesizing complex, multi-hop knowledge.

A trade-off exists: increasing λ enhances diversity by encouraging jumps to new documents but can prematurely terminate paths or fragment reasoning chains. Conversely, a low λ (e.g., $\lambda = 0$) preserves path length but often confines paths to single documents. To determine the optimal configuration, we conducted a grid search for $\lambda \in \{0, 1, 2, 3\}$. The results (Figure 7 and Figure 8) reveal consistent patterns across Linguistics and Legal domains:

- $\lambda = 0$ yields the lowest diversity, with random walks trapped in local clusters.
- $\lambda = 1$ boosts diversity ($\sim$10% improvement) while maintaining healthy long paths (≥ 3 hops).
- $\lambda \geq 2$ sharply declines long path frequency, as the strong bias toward unseen documents forces early termination once few unvisited neighbors remain.

Consequently, we selected $\lambda = 1$ as the optimal setting, balancing rich, multi-source information with deep, multi-step reasoning structures.

E.6 Ablation on KG Quality

To assess the robustness of DKA against potential incompleteness in the knowledge graph, we conducted an experiment by randomly removing a certain percentage of nodes (5%, 10%, 15%) from the graph before synthesizing data for fine-tuning. The results, presented in Table 13, show that performance degradation is minimal across all three domains even with incomplete graphs. For instance, even with a 15% node removal rate, the accuracy drop for Qwen3-8B on Linguistics is less than 1% (from 70.87% to 70.04%), and similar resilience is observed in the Law domain and on MuSiQue (94.66% to 94.38%). This suggests that DKA is robust to graph quality, likely due to the synergistic effect of the weighted random walk and ambiguous node strategies, which encourage the model to learn generalized reasoning patterns rather than overfitting to specific graph structures.

Domain	Model	Removal Rate	Acc Avg
Linguistics	Qwen3-8B	0%	70.87
		5%	70.76
		10%	70.59
		15%	70.04
Law	Qwen3-8B	0%	63.71
		5%	63.68
		10%	63.56
		15%	63.21
Law	Qwen2.5-7B	0%	67.62
		5%	67.58
		10%	67.46
		15%	67.10
MuSiQue	Qwen2.5-7B	0%	94.66
		5%	94.62
		10%	94.54
		15%	94.38

Table 13: Robustness of DKA to Knowledge Graph Incompleteness.

F Case Study

To demonstrate the reasoning capabilities developed through DKA, particularly the efficacy of the "Ambiguous Node Construction" strategy (Section 3.3.3), we present a case study from the Linguistics domain. In this example, the explicit entity "Latin" is replaced by a descriptive definition, requiring the model to first identify the entity and then synthesize the answer. The trace reveals that the model successfully resolves the ambiguity ("ancient language... widespread... Roman Empire" $\rightarrow$ Latin) within its internal reasoning process before formulating the final relationship. This validates that DKA trains the model to perform "reasoning-by-navigation"—traversing from attributes to entity identity and then to related concepts—rather than relying on superficial pattern matching.

Case Study: User Query

What is the relationship between an ancient language of Indo-European origin with widespread distribution throughout the Roman Empire and modern Romance languages?

Case Study: Model Reasoning Process (<think>)

Okay, let's tackle this question. The user is asking about the relationship between an ancient Indo-European language that was widespread in the Roman Empire and modern Romance languages. First, I need to recall the key points. The ancient language in question is Latin, which is part of the Indo-European family. Latin was the official language of the Roman Empire and was used in administration, law, and culture. It's known for its extensive use in the empire, which spans a vast area. Now, modern Romance languages like Spanish, Portuguese, French, Italian, Romanian, and others are derived from Latin. These languages evolved from Latin through processes like phonological changes, syntactic simplification, and lexical borrowing. The key point here is that Latin is the ancestor of these Romance languages. The question mentions the widespread distribution of the ancient language, which aligns with Latin's use across the Roman Empire. The answer should connect Latin's role as the parent language to the Romance languages, explaining how they developed from it. Also, I should structure the reasoning step by step, starting with identifying the ancient language, its distribution, and then its evolution into modern languages. Let me check if there are any other details. I should include that to show the historical context. The answer should clearly state that Latin is the ancestor of the Romance languages, explaining the process of linguistic evolution. I need to ensure that the explanation is clear and concise, without any markdown. Also, the final answer should be stated explicitly, as per the instructions. Let me put this all together now.

Case Study: Model Response

According to my knowledge, the ancient language in question is Latin, which belongs to the Indo-European language family. Latin was the dominant language of the Roman Empire, used for administration, law, education, and cultural expression across its vast territories. Its widespread distribution during the Roman period laid the foundation for the development of modern Romance languages. These languages—such as Spanish, Portuguese, French, Italian, and Romanian—emerged from Latin through a process of linguistic evolution, including phonological shifts, syntactic simplification, and lexical adaptation. Over time, regional variations of Latin diversified into distinct languages, all of which retain structural and lexical ties to their common ancestor. The relationship is thus one of direct ancestry: Latin is the prototypical parent language, while the Romance languages are its descendant languages, sharing a common historical and genetic origin.
Answer: Latin is the ancestor of modern Romance languages, which evolved from it through historical linguistic changes.

G AI Use

We utilized AI assistants to help polish parts of the content in this paper, including correcting grammar and formatting errors, as well as refining wording.