

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

Anonymous ACL submission

Abstract

Achieving deep expertise in vertical domains requires more than exposing Large Language Models (LLMs) to vast corpora; it demands a structured internalization of complex logic. Standard fine-tuning often treats data as a disordered bag of tokens, failing to capture the intricate dependencies essential for high-level reasoning. We propose Domain Knowledge Annealing (DKA), a curriculum learning paradigm designed to maximize domain proficiency. Inspired by physical crystallization, DKA organizes training from Local to Global: it begins by injecting discrete concepts via entity-centric samples ("heating") and progressively advances to synthesizing complex, cross-document relationships ("cooling"). This structured progression allows the model to solidify isolated facts into a coherent knowledge system. Experiments on Linguistics and Law benchmarks demonstrate that DKA significantly surpasses standard strategies, establishing a new state-of-the-art in domain-specific reasoning and understanding.

1 Introduction

The transformative potential of Large Language Models (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 often suffices (Zhao et al., 2023; Minaee et al., 2024; Li et al., 2025), specialized fields demand strict adherence to logical frameworks and the synthesis of 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), which limits their deployment in high-stakes environments.

We attribute this limitation to a misalignment between training dynamics and domain knowledge

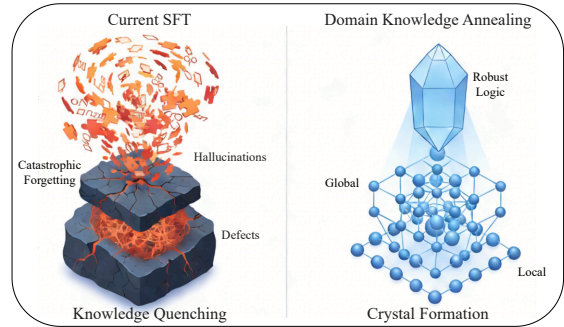


Figure 1: Conceptual comparison of training paradigms. Standard unstructured fine-tuning (Left) acts like "quenching," causing distributional conflicts and hallucinations. In contrast, DKA (Right) adopts an "annealing" schedule—progressing from local facts to global reasoning—to ensure robust knowledge crystallization.

structure. True expertise resembles a crystal lattice, formed by interconnected concepts. In contrast, standard fine-tuning treats data as unstructured "bags of tokens," ignoring logical complexity. Although "Textbooks Are All You Need" (Gunasekar et al., 2023), real-world domain data is often fragmented and unstructured. Traditional data synthesis methods exacerbate this by extracting isolated facts from single documents, yielding simplistic Question-Answer pairs. These local samples fail to challenge the model to integrate disjointed knowledge, resulting in models that are proficient at recalling isolated facts but incompetent at connecting them to solve complex, multi-hop problems.

The lack of a structured curriculum further compounds these issues. Mixing simple factual queries with complex reasoning creates distributional conflicts, akin to "quenching" in metallurgy (Figure 1). This disordered training induces gradient conflicts, hindering the balance between learning atomic facts and high-level reasoning. Moreover, forcing complex reasoning without a solid factual foundation leads to plausible but erroneous reasoning chains (Huang et al., 2024). Additionally, distribution mismatches between synthetic data and pre-

068 training distributions can trigger catastrophic for- 118
069 getting, compromising general capabilities (Kirk- 119
070 patrick et al., 2017; Zhang et al., 2022; Yang et al., 120
071 2024). 121

072 To address these challenges, we introduce Do- 122
073 main Knowledge Annealing (DKA), a structured 123
074 curriculum learning framework designed to crystal- 124
075 lize domain proficiency. Drawing inspiration from 125
076 thermodynamic annealing, DKA organizes training 126
077 through a "Local-to-Global" progression. It begins 127
078 with a "heating" phase for local knowledge injec- 128
079 tion, utilizing single-document data to allow the 129
080 model to absorb high-entropy, discrete concepts. 130
081 This is followed by a controlled "cooling" phase 131
082 that guides the model toward global reasoning inte- 132
083 gration, compelling it to traverse complex Knowl- 133
084 edge Graph (KG) topologies to synthesize informa- 134
085 tion across multiple documents. This curriculum 135
086 mimics the formation of crystal lattices, solidify- 136
087 ing isolated facts into a coherent expert knowledge 137
088 system. 138

089 Methodologically, DKA integrates three core 139
090 components to automate this crystallization pro- 140
091 cess. We first construct a fine-grained domain 141
092 knowledge graph with provenance tracking to map 142
093 the corpus’s logical structure. Leveraging this 143
094 topology, we introduce a data synthesis algorithm 144
095 that utilizes rare node identification and weighted 145
096 random walks to generate cross-document, multi- 146
097 hop QA pairs. Finally, we employ a two-stage 147
098 progressive training regimen that aligns with the 148
099 annealing schedule, minimizing the "alignment tax" 149
100 by using the base model itself for data refinement.

101 Our experiments on Linguistics and Law se- 150
102 lected as representative benchmarks for knowledge- 151
103 intensive vertical domains demonstrate that DKA 152
104 does not merely surpass standard random shuffling 153
105 baselines but achieves state-of-the-art performance, 154
106 often outperforming significantly larger general- 155
107 purpose models. These results reveal a critical find- 156
108 ing: the structured organization of data Specifically, 157
109 the transition from local facts to global reasoning, 158
110 is a decisive factor comparable to data scaling for 159
111 maximizing domain-specific reasoning. 160

112 2 Related Work 162

113 2.1 Data Synthesis 163

114 High-quality Supervised Fine-Tuning data is the 164
115 cornerstone of effective domain adaptation. Yet, 165
116 manual construction of such data—particularly for 166
117 multi-document and multi-hop reasoning—remains 167
168

prohibitively expensive. Early synthesis methods, 118
represented by Self-Instruct (Wang et al., 2023; 119
Jeong, 2024), typically rely on general-purpose 120
priors or single-document extraction. While scal- 121
able, these approaches often yield shallow instruc- 122
tion pairs that fail to challenge a model’s ability 123
to synthesize fragmented knowledge (Lu et al., 124
2025; Susnjak et al., 2025). To address this, recent 125
works have begun to incorporate structural priors. 126
For instance, Synthetic-on-Graph constructs con- 127
text graphs to capture explicit knowledge associa- 128
tions (Ma et al., 2025), and Condor employs World 129
Knowledge Trees with self-reflection to enhance 130
data quality (Cao et al., 2025). Similarly, Wiz- 131
ardLM and Orca focus on progressively increasing 132
instruction complexity or leveraging explanation 133
traces to teach reasoning processes (Xu et al., 2023; 134
Mukherjee et al., 2023). 135

DKA transcends these approaches by transform- 136
ing the Knowledge Graph from a static reference 137
into a dynamic engine for reasoning path genera- 138
tion. Unlike Graph of Thoughts, which structures 139
reasoning only during the inference phase (Besta 140
et al., 2024), DKA internalizes this structure di- 141
rectly into the training data. By integrating rare 142
node identification with weighted random walks, 143
we synthesize complex, multi-hop QA pairs that 144
explicitly require cross-document information syn- 145
thesis. This fills a critical gap in automated data 146
generation, enabling models to learn "reasoning- 147
by-navigation" rather than simple pattern match- 148
ing (Song et al., 2025; Yang et al., 2025b). 149

150 2.2 Staged Training 150

Curriculum learning serves as a vital mechanism 151
for adapting LLMs to specialized domains (Xu 152
et al., 2020). While standard fine-tuning often 153
treats data as a uniform mixture, recent findings 154
indicate that data composition and ordering fun- 155
damentally dictate model capabilities (Dong et al., 156
2024; Liu et al., 2025b). Existing solutions typi- 157
cally rely on heuristic curricula, prioritizing tasks 158
based on surface metrics like sequence length (Kim 159
and Lee, 2024) or model perplexity (Luo et al., 160
2024; Chen et al., 2025). However, these proxies 161
often fail to reflect the true logical density of do- 162
main knowledge; a long sequence is not necessarily 163
logically complex, and high perplexity may stem 164
from noise rather than depth. Consequently, curri- 165
cula based on these shallow metrics often struggle 166
to foster genuine reasoning capabilities, as they do 167
not align with the intrinsic cognitive structure of 168

169 the domain.

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

187 2.3 Annealing

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

203 DKA extends this paradigm by defining a rigorous
204 "Knowledge Annealing" process that is strictly
205 governed by domain topology. We posit that the
206 topological transition from isolated entities to interconnected
207 subgraphs provides the most faithful mapping to the
208 thermodynamic cooling process. The "Heating Phase"
209 corresponds to the injection of high-entropy, local
210 single-document data, allowing the model to absorb
211 discrete, disordered facts. The "Cooling Phase" then
212 transitions to low-entropy, global cross-document
213 reasoning paths, forcing the model to settle these
214 facts into a stable, ordered logical lattice. This
215 structure-based annealing ensures that the "crystallization"
216 of intelligence is not merely a statistical reduction
217 in loss, but a genuine formation of coherent, evidence-
218 grounded reasoning structures.
219

220 Crucially, DKA distinguishes itself from
221 retrieval-based paradigms like GraphRAG. While
222 GraphRAG relies on external knowledge bases to
223 support reasoning during inference (Edge et al.,
224 2024), DKA aims to **internalize** these expert
225 reasoning capabilities directly into the model
226 parameters. By training on topologically structured
227 paths, DKA enables the model to perform expert-
228 level reasoning and synthesis autonomously, without
229 the need for external retrieval. This capability is
230 particularly vital for latency-sensitive or offline
231 scenarios, ensuring that the model functions as a
232 self-sufficient domain expert.

233 3 DKA: Domain Knowledge Annealing

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

237 The design of DKA is motivated by the need to
238 address inherent limitations in learning from
239 unstructured text. Unstructured corpora often obscure
240 logical relationships between distant concepts; to
241 make these explicit, we introduce a Knowledge
242 Graph (KG) that reconstructs such connections.
243 Moreover, since standard training tends to overem-
244 phasize frequent concepts, we leverage rare-node-
245 guided synthesis to ensure exposure to the long
246 tail of knowledge. Drawing an analogy to physical
247 annealing—where high thermal energy allows
248 atoms to rearrange before settling into a low-energy
249 crystalline state—we adopt a progressive training
250 strategy that evolves from local knowledge injection
251 ("heating") to global reasoning crystallization
252 ("cooling"). DKA integrates three core components:
253 (1) domain knowledge graph construction with
254 provenance tracking, (2) automated synthesis of
255 hierarchical SFT data, and (3) a two stage "Local-
256 to-Global" progressive training regimen. The overall
257 workflow is outlined in Algorithm 1.

258 3.1 Domain Knowledge Graph Construction

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

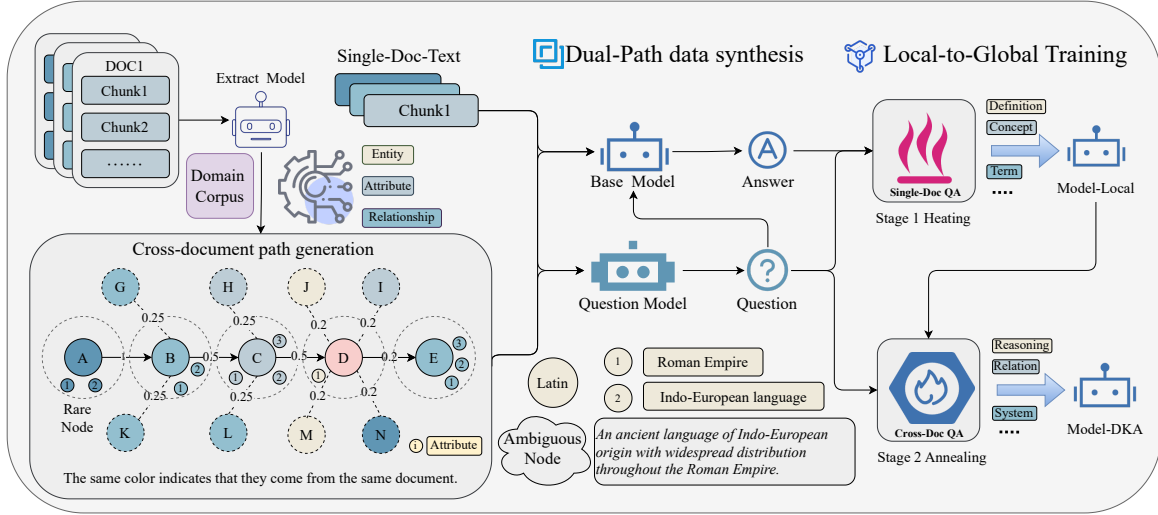


Figure 2: The proposed Domain Knowledge Annealing Framework

Algorithm 1. DKA: Domain Knowledge Annealing

Input: Domain corpus \mathcal{D} , Base LLM M_{base}

Output: Domain-specific LLM M_{final}

Phase 1: Knowledge Graph Construction

- 1: $\mathcal{C} \leftarrow \text{DocumentChunking}(\mathcal{D})$
- 2: $\mathcal{E}, \mathcal{A}, \mathcal{R} \leftarrow \text{InformationExtraction}(\mathcal{C}, M_{\text{extract}})$
- 3: $\mathcal{G} \leftarrow \text{BuildGraph}(\mathcal{E}, \mathcal{A}, \mathcal{R})$
- 4: **for** $v \in \mathcal{V}(\mathcal{G})$ **do**
- 5: $v.\text{source} \leftarrow \text{GetSourceChunk}(v)$

Phase 2: Single-Document Data Synthesis

- 6: $\mathcal{Q}_{\text{single}} \leftarrow \emptyset$
- 7: **for** $c \in \mathcal{C}$ **do**
- 8: $q \leftarrow \text{GenerateQuestion}(M_{\text{question}}, c)$
- 9: $a \leftarrow \text{RefineAnswer}(M_{\text{base}}, q, c)$
- 10: $\mathcal{Q}_{\text{single}}.\text{append}((q, a))$

Phase 3: Cross-Document Data Synthesis

- 11: $\mathcal{V}_{\text{rare}} \leftarrow \text{IdentifyRareNodes}(\mathcal{G})$
- 12: $\mathcal{Q}_{\text{cross}} \leftarrow \emptyset$
- 13: **for** $v_{\text{start}} \in \mathcal{V}_{\text{rare}}$ **do**
- 14: $\mathcal{P} \leftarrow \text{WeightedRandomWalk}(\mathcal{G}, v_{\text{start}})$
- 15: **for** $p \in \mathcal{P}$ **do**
- 16: $p_{\text{amb}} \leftarrow \text{AmbiguifyNode}(p)$
- 17: $q \leftarrow \text{GenerateQuestion}(M_{\text{question}}, p_{\text{amb}})$
- 18: $a \leftarrow \text{RefineAnswer}(M_{\text{base}}, q, p)$
- 19: $\mathcal{Q}_{\text{cross}}.\text{append}((q, a))$

Phase 4: Progressive Training

- 20: $M_{\text{Local}} \leftarrow \text{SFT}(M_{\text{base}}, \mathcal{Q}_{\text{single}})$
- 21: $M_{\text{DKA}} \leftarrow \text{SFT}(M_{\text{Local}}, \mathcal{Q}_{\text{cross}})$
- 22: **return** M_{DKA}

semantically coherent 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 cross-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 Single-Document QA Pair Generation

Single-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 Cross-Document QA Pair Generation

To foster cross-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 Cross-Document Path Generation.

269
270
271
272
273
274

275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306

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 cross-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:

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

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 Cross-Document Path Generation

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

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

where v_t is the node at step t , $D(v_t)$ denotes the document containing node v_t , $\mathcal{D}_t = \{D(v_1), D(v_2), \dots, D(v_t)\}$ 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.6. 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 $P = (v_1, v_2, \dots, v_k)$ 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 cross-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. We introduce ambiguity by masking the explicit entity name and replacing 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. We further discuss and validate the effectiveness of this strategy through ablation studies in Appendix E.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 (Heating). The model is fine-tuned on single-document QA pairs. This phase prioritizes the internalization of domain-specific terminology, definitions, and atomic facts, establishing 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 heating phase

where atoms gain energy and mobility.

Stage 2: Knowledge Annealing (Crystallization). The model is further fine-tuned on cross-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 "cools down" the model, crystallizing 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.

4 Experiments

We evaluate the DKA framework across two distinct domains: Linguistics and Law. 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).

We first analyze the complexity of our synthesized data. Figure 3 illustrates the token length distributions for the Linguistics domain (Legal domain in Appendix E.2). Cross-document samples exhibit significantly longer sequences in both questions and answers compared to single-document samples, reflecting the increased reasoning depth required to synthesize information from disparate sources.

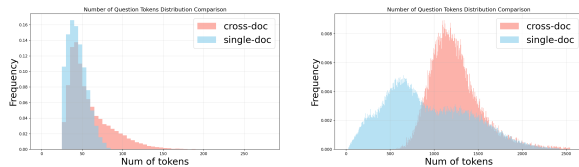


Figure 3: Length distributions of (left) single-document/cross-document questions and (right) their corresponding answers

To benchmark DKA, we compare three distinct training strategies outlined in Table 1: (1) Local FT, trained exclusively on single-document data; (2) Mix FT, trained on a shuffled mixture of single- and cross-document data; and (3) DKA FT, our proposed progressive approach that evolves from

| Strategy | Single-Doc | Cross-Doc | Local-to-Global |
|----------|------------|-----------|-----------------|
| Local FT | ✓ | × | × |
| Mix FT | ✓ | ✓ | × |
| DKA FT | ✓ | ✓ | ✓ |

Table 1: Comparison of different training strategies

local knowledge injection to global reasoning crystallization.

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, 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. Among the evaluated models, DeepSeek-V3.2Exp achieves the best performance of 71.76%, attributed to its massive 671B parameters. 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 61.22%, Single Choice 45.36%). 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

| 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 %).

(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 (53.02%). Remarkably, applying DKA yielded a dramatic improvement, boosting the score to 63.71%—an increase of over 10%. This demonstrates that DKA is not limited to refining already strong models but can also significantly enhance models with weaker initial domain capabilities. These experiments reveal a critical insight: Regardless of the base model’s initial capability, 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 single-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 cross-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%). This validates our hypothesis that structured, progressive training is essential for achieving expert-level performance and confirms that DKA is a robust, model-agnostic framework that systematically cultivates expert-level reasoning capabilities.

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 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.

This phenomenon aligns with the thermodynamic principles of annealing. DKA FT effectively decouples the learning objectives: the first stage acts as a "heating" phase, injecting high-entropy factual knowledge, while the second stage serves as a controlled "cooling" phase. In this latter stage, the constraints of complex cross-document reasoning force the model to organize this knowledge into a stable, low-energy configuration—analogueous to crystallization. This separation prevents the "quenching" effect often seen in Mix FT (where gradient conflicts trap the model in suboptimal lo-

| 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 | 54.75 | 25.72 | 43.33 | 55.17 | 53.02 |
| + 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

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

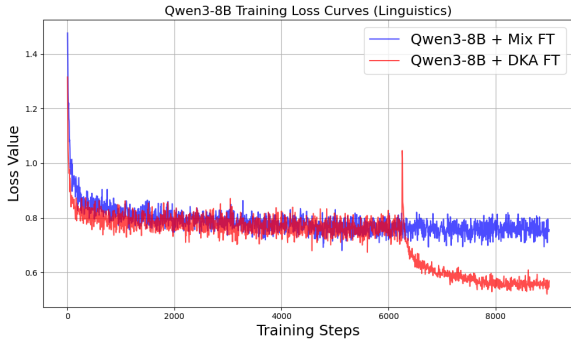


Figure 4: Comparison of training loss curves based on Qwen3-8B (Linguistics). The DKA FT strategy converges to a lower loss than Mix FT strategy.

cal minima) and facilitates the formation of a robust expert-level representation.

4.3 Impact on General Capabilities

We evaluated the models on MMLU (Hendrycks et al., 2020), superGPQA (Team, 2025), and BBH (Suzgun et al., 2023) to assess the "alignment tax". Results in Table 4 show that degradation is minimal. For instance, the MMLU score (75.55%) of Qwen3-8B + DKA FT(Lingua) is comparable to the base model (76.89%). Similarly, in the legal domain, Qwen3-8B + DKA FT(Law) maintains a high BBH score of 75.93%, only slightly below the baseline of 78.40%. This stability is attributed to our self-generation strategy, which maintains the data distribution close to the model's pre-training manifold. By enabling the base model to generate answers based on the provided reference text, we ensure that the "style" of the SFT data matches

| Model | MMLU | superGPQA | BBH |
|-------------------|-------|-----------|-------|
| Qwen3-8B | 76.89 | 31.64 | 78.40 |
| + DKA FT (Lingua) | 75.55 | 29.88 | 75.56 |
| + DKA FT (Law) | 74.54 | 29.74 | 75.93 |
| Qwen2.5-7B | 74.16 | 26.34 | 70.04 |
| + DKA FT (Law) | 70.65 | 25.96 | 66.95 |

Table 4: Performance on general benchmarks. DKA preserves general capabilities.

the model's internal representation, preventing the "catastrophic forgetting" often caused by training on distributionally distinct text.

5 Conclusion

In this work, we introduced Domain Knowledge Annealing (DKA), a structured paradigm that transforms domain adaptation from simple data exposure into rigorous knowledge crystallization. By integrating a provenance-aware knowledge graph with a "Local-to-Global" curriculum, DKA systematically guides models from absorbing atomic facts to synthesizing stable reasoning structures, mirroring thermodynamic annealing. Our evaluations across Linguistics and Law—including the novel Lingua-Bench—demonstrate that DKA achieves state-of-the-art performance, often outperforming significantly larger general-purpose models. These findings reveal a fundamental insight: the structural organization of training data is as decisive as data quantity, paving the way for the efficient creation of trustworthy, high-precision expert LLMs in high-stakes domains.

599 Limitations

600 While DKA demonstrates significant improve-
601 ments in domain adaptation, it is not without lim-
602 itations. First, the quality of our synthetic data is
603 intrinsically tied to the accuracy of the underlying
604 corpus. Errors or biases in the source text can prop-
605 agate into the knowledge graph and subsequently
606 into the training data. However, our system’s prove-
607 nance tracking capability (mapping graph nodes
608 back to source text chunks) provides a mechanism
609 to identify and rectify such errors at the source,
610 partially mitigating this risk. Second, compared to
611 standard SFT data synthesis, DKA introduces an
612 additional overhead for Knowledge Graph construc-
613 tion. We have optimized this process for efficiency
614 (as detailed in Appendix E.2). Finally, while our
615 progressive training strategy significantly reduces
616 catastrophic forgetting compared to random mix-
617 ing, a marginal trade-off with general capabilities
618 persists. Future work will focus on developing
619 dynamic graph update mechanisms and exploring
620 more robust regularization techniques to further
621 narrow this gap.

622 Ethical Considerations

623 The development of domain-specific large lan-
624 guage models carries inherent ethical responsibili-
625 ties. We acknowledge that our models, particularly
626 legal models, are intended for research and assistive
627 purposes only and should not be used as a substi-
628 tute for professional legal advice. The potential
629 for bias in the training data—stemming from the
630 original domain corpora—remains a concern. We
631 have made efforts to filter harmful content during
632 the data synthesis phase, but we cannot guarantee
633 the complete absence of biases. Furthermore, the
634 automated generation of reasoning chains, while
635 powerful, could potentially be misused to generate
636 plausible-sounding but factually incorrect misinfor-
637 mation if not carefully monitored. We urge users
638 to deploy these models with appropriate human
639 oversight and safety guardrails.

640 References

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

646 Maciej Besta, Nils Blach, Ales Kubicek, Robert Ger-
647 stenberger, Michal Podstawski, Lukas Gianinazzi,

Joanna Gajda, Tomasz Lehmann, Hubert Niewiadow-
ski, Piotr Nyczyk, and Torsten Hoeffler. 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 Kamaloo. 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 e-prints*, pages arXiv–2407.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitanaky, 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.

| | | | | |
|-----|---|--|--|--|
| 705 | Textbooks are all you need. <i>arXiv preprint arXiv:2306.11644</i> . | | | |
| 706 | | | | |
| 707 | Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data . In <i>Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)</i> , pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics. | | | |
| 708 | | | | |
| 709 | | | | |
| 710 | | | | |
| 711 | | | | |
| 712 | | | | |
| 713 | | | | |
| 714 | | | | |
| 715 | | | | |
| 716 | Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding . <i>CoRR</i> , abs/2009.03300. | | | |
| 717 | | | | |
| 718 | | | | |
| 719 | | | | |
| 720 | 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 . <i>arXiv preprint arXiv:2404.06395</i> . | | | |
| 721 | | | | |
| 722 | | | | |
| 723 | | | | |
| 724 | | | | |
| 725 | | | | |
| 726 | 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 <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1416–1428, Bangkok, Thailand. Association for Computational Linguistics. | | | |
| 727 | | | | |
| 728 | | | | |
| 729 | | | | |
| 730 | | | | |
| 731 | | | | |
| 732 | | | | |
| 733 | | | | |
| 734 | Cheonsu Jeong. 2024. Fine-tuning and utilization methods of domain-specific llms . <i>CoRR</i> , abs/2401.02981. | | | |
| 735 | | | | |
| 736 | Jisu Kim and Juhwan Lee. 2024. Strategic data ordering: Enhancing large language model performance through curriculum learning . <i>arXiv preprint arXiv:2405.07490</i> . | | | |
| 737 | | | | |
| 738 | | | | |
| 739 | | | | |
| 740 | James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks . <i>Proceedings of the National Academy of Sciences</i> , 114(13):3521–3526. | | | |
| 741 | | | | |
| 742 | | | | |
| 743 | | | | |
| 744 | 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 <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 5131–5157, Vienna, Austria. Association for Computational Linguistics. | | | |
| 745 | | | | |
| 746 | | | | |
| 747 | | | | |
| 748 | | | | |
| 749 | | | | |
| 750 | | | | |
| 751 | | | | |
| 752 | Yucheng Li, Shuai Wang, Hongzhou Ding, and 1 others. 2023. Large language models in finance: A survey . In <i>Proceedings of the Fourth ACM International Conference on AI in Finance</i> , pages 374–382. | | | |
| 753 | | | | |
| 754 | | | | |
| 755 | | | | |
| 756 | 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 . <i>arXiv preprint arXiv:2512.02556</i> . | | | |
| 757 | | | | |
| 758 | | | | |
| 759 | | | | |
| 760 | | | | |
| | | 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 <i>Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 29443–29464, Vienna, Austria. Association for Computational Linguistics. | | 761 762 763 764 765 766 767 768 |
| | | Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts . <i>arXiv preprint arXiv:1608.03983</i> . | | 769 770 771 |
| | | Rui Lu, Zhenyu Hou, Zihan Wang, and 1 others. 2025. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl . <i>arXiv preprint, arXiv:2509.10446</i> . | | 772 773 774 775 |
| | | 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 <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics. | | 776 777 778 779 780 781 782 |
| | | Yunfei Luo, Yuyang Liu, Rukai Cai, and Tauhidur Rahman. 2024. Start simple: Progressive difficulty multi-task learning . In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)</i> , pages 48–55, Mexico City, Mexico. Association for Computational Linguistics. | | 783 784 785 786 787 788 789 790 |
| | | Shengjie Ma, Xuhui Jiang, Chengjin Xu, and 1 others. 2025. Synthesize-on-graph: Knowledgeable synthetic data generation for continue pre-training of large language models . <i>arXiv preprint, arXiv:2505.00979</i> . | | 791 792 793 794 795 |
| | | Shervin Minaee, Tomas Mikolov, Nal Nikzad, and 1 others. 2024. Large language models: A survey . <i>arXiv preprint, arXiv:2402.06196</i> . | | 796 797 798 |
| | | Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4 . <i>arXiv preprint arXiv:2306.02707</i> . | | 799 800 801 802 803 |
| | | Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap . <i>IEEE Transactions on Knowledge and Data Engineering</i> , 36(7):3580–3599. | | 804 805 806 807 808 |
| | | 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 <i>Findings of the Association for Computational Linguistics: EMNLP 2025</i> , pages 25297–25311, Suzhou, China. Association for Computational Linguistics. | | 809 810 811 812 813 814 815 |

| | | | |
|-----|--|--|--|
| 816 | 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 <i>The Twelfth International Conference on Learning Representations</i> . | | |
| 817 | | | |
| 818 | | | |
| 819 | | | |
| 820 | | | |
| 821 | | | |
| 822 | Zihan Sun. 2023. A short survey of viewing large language models in legal aspect. <i>arXiv preprint</i> , arXiv:2303.09136. | | |
| 823 | | | |
| 824 | | | |
| 825 | T. Susnjak, P. Hwang, N. Reyes, and 1 others. 2025. Automating research synthesis with domain-specific large language model fine-tuning. <i>ACM Transactions on Knowledge Discovery from Data</i> , 19(3):1–39. | | |
| 826 | | | |
| 827 | | | |
| 828 | | | |
| 829 | 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 <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 13003–13051, Toronto, Canada. Association for Computational Linguistics. | | |
| 830 | | | |
| 831 | | | |
| 832 | | | |
| 833 | | | |
| 834 | | | |
| 835 | | | |
| 836 | | | |
| 837 | M-A-P Team. 2025. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines . <i>CoRR</i> , abs/2502.14739. | | |
| 838 | | | |
| 839 | | | |
| 840 | Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kabilan Elangovan, and 1 others. 2023. Large language models in medicine. <i>Nature Medicine</i> , 29(8):1930–1940. | | |
| 841 | | | |
| 842 | | | |
| 843 | | | |
| 844 | 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 . <i>ArXiv</i> , abs/2409.18839. | | |
| 845 | | | |
| 846 | | | |
| 847 | | | |
| 848 | | | |
| 849 | | | |
| 850 | | | |
| 851 | 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 <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 13484–13508. | | |
| 852 | | | |
| 853 | | | |
| 854 | | | |
| 855 | | | |
| 856 | | | |
| 857 | | | |
| 858 | Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 6095–6104, Online. Association for Computational Linguistics. | | |
| 859 | | | |
| 860 | | | |
| 861 | | | |
| 862 | | | |
| 863 | | | |
| 864 | | | |
| 865 | 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. <i>arXiv preprint arXiv:2304.12244</i> . | | |
| 866 | | | |
| 867 | | | |
| 868 | | | |
| 869 | | | |
| 870 | An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang | | |
| 871 | | | |
| | | Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. <i>arXiv preprint arXiv:2505.09388</i> . | 872 873 874 |
| | | 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 <i>Findings of the Association for Computational Linguistics: ACL 2025</i> , pages 3236–3256, Vienna, Austria. Association for Computational Linguistics. | 875 876 877 878 879 880 881 |
| | | 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 <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics. | 882 883 884 885 886 887 888 889 |
| | | 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 . <i>Preprint</i> , arXiv:2309.11325. | 890 891 892 893 894 895 |
| | | T. Zhang, X. Wang, B. Liang, and 1 others. 2022. Catastrophic interference in reinforcement learning: A solution based on context division and knowledge distillation. <i>IEEE Transactions on Neural Networks and Learning Systems</i> , 34(12):9925–9939. | 896 897 898 899 900 |
| | | Wayne Xin Zhao, Kun Zhou, Junyi Li, and 1 others. 2023. A survey of large language models. <i>arXiv preprint</i> , arXiv:2303.18223. | 901 902 903 |
| | | Tianshi Zheng, Zheyue Deng, Hong Ting Tsang, Weiqi Wang, Jiabin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery . In <i>Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing</i> , pages 17744–17761, Suzhou, China. Association for Computational Linguistics. | 904 905 906 907 908 909 910 911 |

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.

B Benchmark Statistics

B.1 Lingua-Bench Statistics

Table 5 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 (979 items) are curated from graduate entrance examinations, assessing fundamental factual recall and concept definition. "Challenging" questions (820 items) are synthesized via a cross-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 | 637 |
| Single Choice | Easy | 263 |
| Single Choice | Challenging | 183 |
| Total | | 1799 |

Table 5: 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 6, the benchmark encompasses a wide spectrum of legal expertise, ranging from fundamental concepts (e.g., Legal Basics Knowledge) to highly specialized professional qualifications (e.g., National Judicial Exam, Patent Agent Exam). 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 6: 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

989 domain-specific optimization does not compromise
990 the model’s fundamental reasoning and general
991 knowledge base.

992 C Knowledge Graph Extraction

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

1000 For the entity type, the extracted information in-
1001 cludes the entity name, the source document where
1002 the entity is located, and the relevant attributes of
1003 the entity. For the relation type, the extracted con-
1004 tent covers the relation name, the name of the head
1005 entity involved in the relation, the name of the tail
1006 entity, and the corresponding source document of
1007 the relation.

1008 The primary reason for adopting LangExtract
1009 lies in its document traceability capability: it can
1010 label the source document of each extracted en-
1011 tity and relation, which lays a solid foundation
1012 for cross-document weighted random walk. The
1013 extracted structured data can be converted with
1014 minimal effort and then imported into Neo4j, on
1015 which our knowledge graph is built. Subsequent
1016 tasks, including rare node identification and cross-
1017 document weighted random walk, are also imple-
1018 mented by leveraging the GraphDatabase module
1019 of the Python Neo4j library.

1020 D Prompt Templates

1021 D.1 Single-Document Data Synthesis

Question Generation Prompt

Description: You are an expert in [Domain]
text analysis and question design, capable of ex-
tracting key [Domain] information from complex
paragraphs and generating high-quality ques-
tions 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 accu-
rate answers to questions from the given content,
ensuring the accuracy and relevance of the an-
swers.

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.

1023
1024 This section details the prompt templates em-
1025 ployed in our DKA framework for automated data
1026 synthesis. These prompts are designed to guide
1027 the base model in generating high-quality, domain-
1028 specific Q&A pairs. We categorize them into two
1029 distinct phases: Single-Document Data Synthe-
1030 sis, which focuses on extracting atomic knowledge
1031 from individual texts, and Cross-Document Data
1032 Synthesis, which leverages the Knowledge Graph
1033 to construct complex, multi-hop reasoning tasks.

D.2 Cross-Document Data Synthesis 1034

Question Generation Prompt

You are a senior expert in [Domain], tasked
with designing a challenging multi-hop reason-
ing question for an advanced [Domain] course.
Based on the following knowledge graph infor-
mation, craft a question centered around the
"rare_node" that requires at least three reason-
ing 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:** $4 \times 32\text{GB}$ 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 single-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, generating 44,006 cross-document paths. Similarly, for the Legal domain (180,918 nodes), we combined 60,000 single-document QAs with 23,031 cross-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 graph required 14 hours (Table 7). 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 8.

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

Table 7: Statistics of Knowledge Graph Construction

| Domain | Single-Document QA | Cross-Document QA | Total |
|-------------|--------------------|-------------------|---------|
| Linguistics | 100,000 | 44,006 | 144,006 |
| Legal | 60,000 | 23,031 | 83,031 |

Table 8: Scale of the Supervised Fine-Tuning (SFT) Datasets

E.3 QA Token Distribution Analysis

The length distribution of questions and answers in the legal domain QA dataset is shown in Figure 5. The question length distribution (left) shows a diverse range of query complexities, while the answer length distribution (right) confirms that our model is trained to generate comprehensive, detailed responses rather than short, extractive answers. This distribution 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 6 and 7 show the training loss curves for the legal domain based on Qwen3

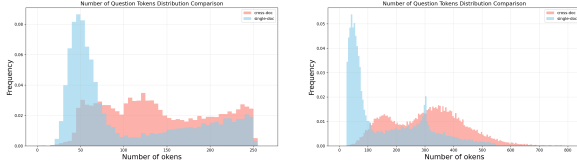


Figure 5: Length distributions of (left) single-document/cross-document questions and (right) their corresponding answers in the legal domain

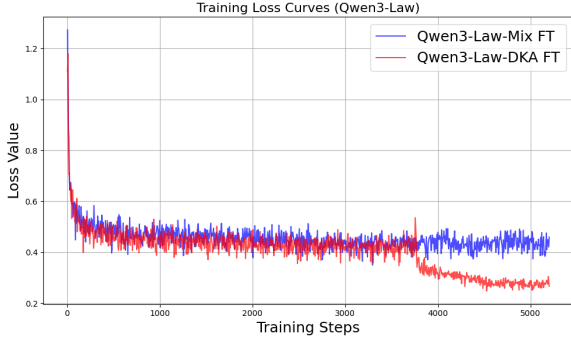


Figure 6: Comparison of training loss curves based on Qwen3-8B (Law). The DKA FT strategy converges to a lower loss than Mix FT.

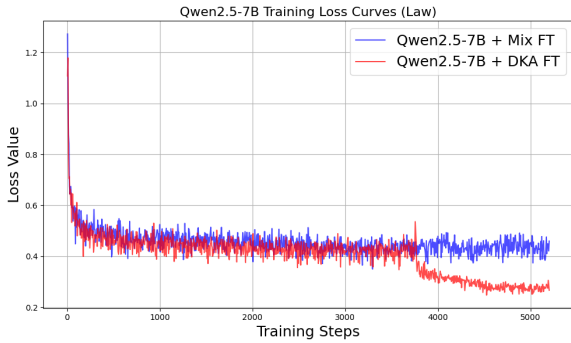


Figure 7: Comparison of training loss curves based on Qwen2.5-7B (Law). The DKA FT strategy converges to a lower loss than Mix FT.

and Qwen2.5, respectively. 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" (single-document) stage to the "global" (cross-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.

E.5 Ablation on Ambiguous Nodes

To validate the contribution of the "Ambiguous Node" strategy (Section 3.3.3) in fostering deep rea-

| Domain | Model | Ambiguous Node | Acc Avg |
|-------------|------------|----------------|---------|
| Linguistics | Qwen3-8B | ✓ | 70.87 |
| Linguistics | Qwen3-8B | ✗ | 69.53 |
| Law | Qwen3-8B | ✓ | 63.71 |
| Law | Qwen3-8B | ✗ | 63.23 |
| Law | Qwen2.5-7B | ✓ | 67.62 |
| Law | Qwen2.5-7B | ✗ | 66.94 |

Table 9: Ablation on Ambiguous Node

soning, we conducted an ablation study. We compared the standard DKA training against a variant where the second stage uses explicit entity names in the cross-document QA pairs. Table 9 presents the results for both the Linguistics and Legal domains. The results show that the model trained with ambiguous nodes consistently outperforms the one trained without them across all settings. This suggests that without ambiguity, the model may partially rely on shortcut learning (e.g., entity name matching) rather than internalizing the full reasoning path. The ambiguous node strategy effectively forces the model to perform "reasoning-by-navigation," thereby enhancing its robustness and depth of understanding.

E.6 Hyperparameter Determination

The parameter λ serves as a regularization term in our weighted random walk algorithm, controlling the bias towards exploring new documents versus staying within the current context. Its value critically influences the structural properties of the generated reasoning paths, specifically affecting both the path length distribution and the semantic richness (document diversity).

We define **Document Diversity** to quantify the information breadth of a reasoning path. Formally, it is calculated as:

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

where N denotes the total number of reasoning paths, $|P_i|$ is the length of path P_i , and $|\{\mathcal{D}(v) \mid v \in P_i\}|$ represents the number of distinct documents associated with nodes in path P_i . A higher diversity score indicates that the reasoning chain aggregates information from a wider array of source documents, which is essential for synthesizing complex, multi-hop knowledge.

However, a trade-off exists. Increasing λ encourages the random walker to jump to nodes in different documents, thereby enhancing diversity. Yet, excessively high values can prematurely ter-

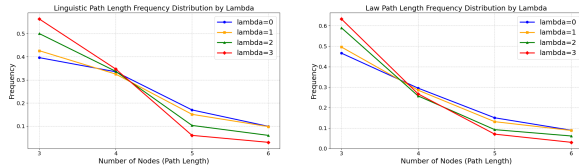


Figure 8: The effect of different λ values on link length distribution for Linguistics (left) and Legal (right) domains.

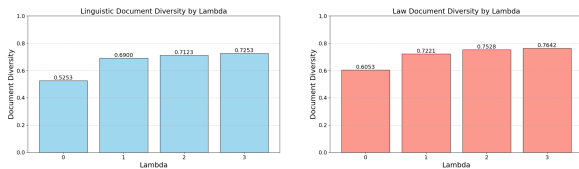


Figure 9: The effect of different λ values on document diversity for Linguistics (left) and Legal (right) domains.

minate paths or fragment the reasoning chain, reducing the number of long, coherent paths necessary for deep reasoning training. Conversely, a low λ (e.g., $\lambda = 0$, corresponding to a pure random walk) preserves path length but often results in paths confined to a single document, limiting the model’s exposure to cross-document inference. To determine the optimal configuration, we conducted a grid search for $\lambda \in \{0, 1, 2, 3\}$. The results, illustrated in Figure 8 and Figure 9, reveal consistent patterns across both Linguistics and Legal domains. Specifically:

- $\lambda = 0$ yields the lowest diversity, indicating that pure random walks tend to get trapped in local document clusters.
- $\lambda = 1$ provides a significant boost in diversity (approximately 10% improvement) while maintaining a healthy distribution of long paths (path lengths ≥ 3).
- $\lambda \geq 2$ leads to a sharp decline in the frequency of long paths, as the strong penalty for revisiting documents forces the walker to terminate early or loop ineffectively.

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

F Case Study

Case Study: Ambiguous Entity Reasoning

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?

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.

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.

To demonstrate the reasoning capabilities developed through DKA, particularly the efficacy of the

1189 "Ambiguous Node Construction" strategy (Section
1190 3.3.3), we present a case study from the Linguistics
1191 domain. In this example, the explicit entity
1192 "Latin" is replaced by a descriptive definition, re-
1193 quiring the model to first identify the entity and
1194 then synthesize the answer. The trace reveals that
1195 the model successfully resolves the ambiguity ("an-
1196 cient language... widespread... Roman Empire" →
1197 Latin) within its internal reasoning process before
1198 formulating the final relationship. This validates
1199 that DKA trains the model to perform "reasoning-
1200 by-navigation"—traversing from attributes to entity
1201 identity and then to related concepts—rather than
1202 relying on superficial pattern matching.

1203 **G AI Use**

1204 We utilized AI assistants to help polish parts of the
1205 content in this paper, including correcting grammar
1206 and formatting errors, as well as refining wording.