Bottom-up Domain-specific Superintelligence: A Reliable Knowledge Graph is What We Need

Anonymous authors
Paper under double-blind review

Abstract

Language models traditionally utilized for cross-domain generalization in natural language understanding and generation have recently demonstrated task-specific reasoning through inference-time scaling. However, their top-down training approach on general text corpora is insufficient for acquiring domain-specific abstractions required for deep expertise in a particular domain. This may require a bottom-up approach that acquires deep expertise by explicitly learning to compose simple concepts of a domain into more complex ones. A knowledge graph (KG) provides such an abstraction where domain primitives are captured by head-relation-tail triples. A KG path formed by such triples captures a higher-level concept. We present a task generation pipeline that directly synthesizes tasks from the domain-specific primitives, enabling the model to explicitly acquire and compose these primitives for reasoning. We fine-tune language models on the resultant bottom-up KG-grounded curriculum to demonstrate domain-specific superintelligence.

Although our approach is readily applicable to a wide variety of domains, we validate it in the context of medicine where reliable KGs are available. Applying our proposed pipeline to a medical KG, we curate a dataset of 24,000 high-quality reasoning tasks paired with structured thinking traces derived from diverse medical primitives. We fine-tune the QwQ-32B model on this bottom-up curriculum to obtain QwQ-Med-3 that takes a step towards medical superintelligence. We also introduce an evaluation suite, ICD-Bench, to quantify domain-specific capabilities of models on reasoning tasks across 15 medical domains. Our experiments demonstrate that QwQ-Med-3 significantly outperforms state-of-the-art open-source and proprietary reasoning models on all categories of ICD-Bench. Further analysis reveals that QwQ-Med-3 utilizes acquired primitives to especially widen the performance gap on the hardest tasks in ICD-Bench. Finally, evaluation on external medical question-answer benchmarks shows that QwQ-Med-3 is able to transfer acquired expertise to improve the performance of the base model.

The industry's approach to artificial general intelligence (AGI) centers on breadth of acquired expertise. We envision a future in which a compositional model of AGI emerges from interacting superintelligent agents, much like how the human society hierarchically acquires ever deeper expertise by combining the expertise of a group of individuals in adjacent domains or super-domains. Furthermore, since language models that are fine-tuned for superintelligence can be relatively small (e.g., 32B parameters), this bottom-up approach may also significantly cut down on training/inference energy costs.

1 Introduction

Recent advances in language modeling (OpenAI, 2024; Anthropic, 2025; Groeneveld et al., 2024; Team et al., 2024; DeepSeek-AI, 2024; Jiang et al., 2024; Yang et al., 2025a; Meta AI, 2025) have made a significant stride towards a cognitive system (Sumers et al., 2023; Park et al., 2023b) capable of performing a wide spectrum of tasks with human-like proficiency (Anthropic, 2025; OpenAI, 2025a;b; Google Gemini Team, 2024). Yet, human-level generality may only be a waypoint on the path to advanced intelligent systems that may exceed

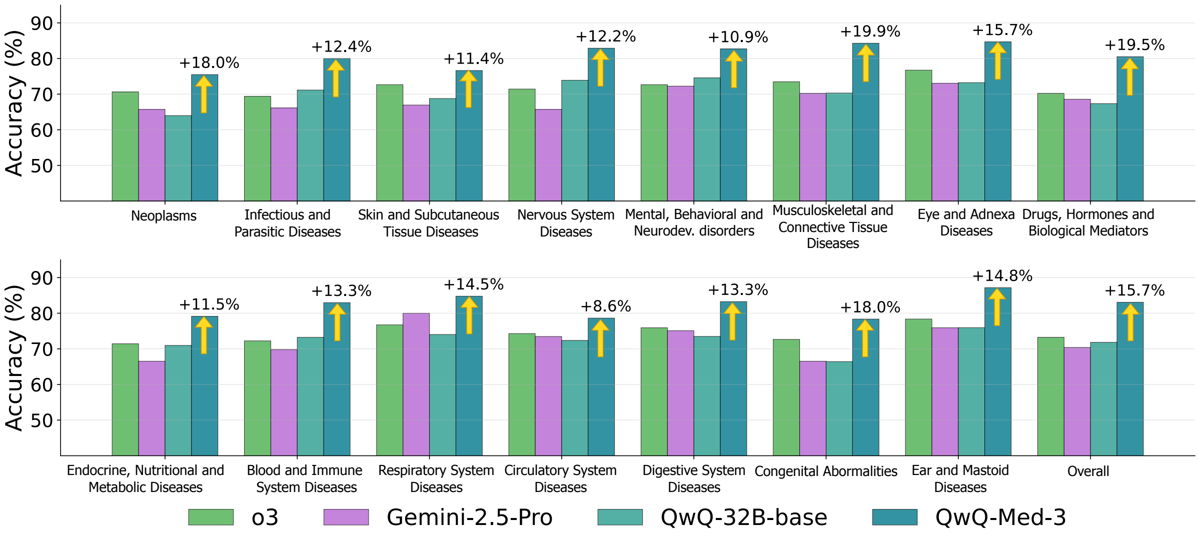


Figure 1: We traverse paths on a medical KG to generate 24,000 tasks that abstract diverse medical primitives. Our QwQ-Med-3 model, fine-tuned on these curated tasks, elicits domain-specific reasoning abilities that significantly outperform other reasoning models across medical sub-specialties.

the cognitive performance of humans: Superintelligence (Bostrom, 2014; Burns et al., 2023). While achieving the breadth of human cognition is one goal of advanced artificial intelligence, superintelligence might be orthogonally characterized by depth, outperforming the best human experts in specialized domains (Jumper et al., 2021; Mankowitz et al., 2023; Novikov et al., 2025; Google DeepMind, 2025a; Gottweis et al., 2025; Merchant et al., 2023; Brixi et al., 2025), like proving unsolved conjectures in number theory, developing novel kinase inhibitors for rare cancer subtypes, or discovering new ferromagnetic semiconductors that operate at room temperature. Superintelligence is meaningful only in the context of a specific domain, where depth of capability is essential and can be concretely characterized. Consequently, advancing towards 'domain-specific' superintelligence might require fine-tuning general cross-domain intelligence into specialized domain-specific expertise.

Language models (LMs) have achieved remarkable cross-domain performance in natural language understanding and generation, largely through scaled pre-training (Hoffmann et al., 2022; Brown et al., 2020). Most recently, scaling inference-time compute (Snell et al., 2024; Gao et al., 2024b; Muennighoff et al., 2025) in pre-trained models via reinforcement learning (DeepSeekAI, 2025) or post-training on high-quality data (Lambert et al., 2024) has been shown to elicit deeper task-specific reasoning. The emergent capacity for specialized reasoning within generalist LMs may suggest that they could serve as a foundation for scaling toward superintelligent specialists (Wei et al., 2022). However, these models are fundamentally limited by their *top-down* approach to learning: They acquire general abstractions of the world through self-supervised learning on vast datasets that may predominantly capture surface-level regularities of a domain (Schaeffer et al., 2023; Turpin et al., 2023; Garg et al., 2023; Liu et al., 2023; Qi et al., 2024; Chollet et al., 2025). In contrast, bottom-up learning builds knowledge hierarchically from first principles rather than inferring structure from aggregated patterns. Acquiring deep expertise in a field necessitates such a *bottom-up* understanding, starting with axioms that capture fundamental relationships among concepts of the domain and then composing them to build upwards to a higher-order understanding (Carey, 2009; Tenenbaum et al., 2011; Lake et al., 2016). This kind of bottom-up organization is difficult to find and acquire through Internet-derived general text corpora. For example, a student builds expertise by following the pedagogical structure of a textbook, beginning with foundational chapters and gradually progressing to more advanced chapters, not merely reading encyclopedic summaries. Past pioneering works, in neurosymbolic reasoning (Ellis et al., 2020) and probabilistic graph inference (Zhang et al., 2020), have attempted to develop hierarchical domain expertise from primitives but have failed to generalize beyond synthetic regimes. Conversely, LMs demonstrate incredible generalizability but lack grounding in structured knowledge. This motivates the central question of our work:

Can explicitly training LMs on structured domain knowledge via a bottom-up curriculum elicit the emergence (if any) of a domain-specific superintelligence?

Naturally, the question then arises: How do we organize domain knowledge into a structured curriculum from which an LM can effectively learn? Knowledge graphs (KGs) (Ji et al., 2022) offer a useful scaffold for structuring knowledge that can tackle this challenge. KGs organize information as a rich graph database where nodes represent semantically meaningful entities from the domain and edges denote the relationships between them. Each edge typically captures a primitive relation in the form of a (head entity, relation, tail entity) triple. For example, `(Methane, Contains Element, Carbon)` represents the axiomatic fact that methane molecules contain carbon atoms. Edges further facilitate composite relational reasoning through the traversal of multi-hop paths along a chain of interconnected edges. For example, `(Methane, Contains Bond, C-H Bond), (C-H Bond, Is Type Of, Sigma Bond), (Sigma Bond, Has Property, Single Covalent Bond)` captures the bonding structure of methane, where `C-H` bonds are sigma bonds that possess the property of being single covalent bonds. A KG comprises many such paths whose local topology naturally induces a bottom-up curriculum, beginning with atomic relations and composing them into more complex reasoning chains.

Our core insight is that paths in a KG can be translated into grounded natural language reasoning tasks, whose solution requires reasoning along the relational chain encoded in the paths. Training on such tasks can then enable an LM to explicitly acquire domain primitives and learn how to systematically compose them at inference time. To this end, we design a task-generation pipeline that can distill high-quality reasoning tasks into a 'curriculum' grounded in the KG paths. More specifically, we use a backend large language model (LLM) to map a KG path into a closed-ended question-answering (QA) task. In addition to the QA pair, we distill detailed, step-by-step thinking traces from a reasoning LLM to elucidate the relational chain on the KG path. Note that curriculum generation with a reasoning LLM incurs only a one-time cost. Generating multiple such tasks across the KG and fine-tuning an LM on them yields a *curriculum-tuned* model that can effectively elicit deeper reasoning (see Fig. 1) by composing acquired bottom-up KG primitives.

While our proposed approach is domain-agnostic, given a reliable KG, we demonstrate it in the context of medicine, where such a KG is readily available. Medical experts have meticulously curated the Unified Medical Language System (UMLS) KG (Bodenreider, 2004; Yasunaga et al., 2021), which offers a rich, structured repository of medical concepts and their interrelations. This makes medicine an ideal testbed for our approach: Deriving a curriculum from the UMLS KG paths enables reasoning abilities beyond recalling isolated medical facts to traversing the complex web of diseases, drugs, symptoms, and underlying biological mechanisms. Models fine-tuned on UMLS-grounded curricula should be able to effectively compose learned KG primitives for reasoning across medical sub-specialties, where each domain requires deep, expert-level understanding to interpret complex clinical vignettes. To this end, we introduce ICD-Bench, an evaluation suite comprising medical QA tasks stratified across 15 categories of the International Classification of Diseases (ICD) taxonomy (Organization, 1992). Each ICD-Bench task is grounded in novel paths composed of domain-specific nodes from the UMLS KG, requiring the reasoning chain to correctly recall and traverse KG primitives along the path. Thus, ICD-Bench provides a reliable probe for bottom-up domain-specific superintelligence by evaluating whether models demonstrate scalable, compositional reasoning across specialized clinical domains.

Our experiments on ICD-Bench reveal the emergence of domain-specific reasoning in curriculum-tuned models that significantly outperform strong baselines, including proprietary and open-source reasoning models, across all 15 categories of ICD-Bench (Fig. 1). We demonstrate that performance improves with deeper and more diverse KG curricula, with curriculum depth proving especially crucial for the most challenging reasoning tasks. Fine-grained ablations further disentangle the contributions of path length, diversity, and complexity sampling, revealing that compute-optimal curricula depth distribution should adapt to task difficulty. Finally, we demonstrate that curriculum-tuned models can transfer acquired KG primitives to external medical QA benchmarks beyond the original KG. Concretely, our contributions are twofold:

1. We propose generating a bottom-up curriculum for LMs using a reliable domain-specific KG (Section 3). Our task-generation pipeline (Section 3.1) traverses multi-hop paths in the KG to produce natural language QA

tasks grounded in domain primitives (Section 3.1.1). Each QA task is paired with a high-quality thinking trace derived from its underlying KG path, enabling the construction of a training curriculum (Section 3.1.2). We instantiate this framework on the expert-curated UMLS medical KG, generating 24,000 QA tasks and associated thinking traces (Section 5.1). We fine-tune the off-the-shelf QwQ reasoning model on this curriculum, resulting in a curriculum-tuned model with acquired bottom-up KG primitives (Section 3.2).

2. We introduce the ICD-Bench evaluation suite (Section 4), comprising domain-specific QA pairs across medical sub-specialties. Our experiments demonstrate that our curriculum-tuned models significantly outperform other reasoning baselines across the ICD-Bench categories (Section 6.1), improve performance robustness to challenging tasks (Section 6.3), and can be optimally adapted to varying task difficulties (Section 6.4). They also demonstrate effective composition of KG primitives (Section 6.5) and transferability to other medical QA benchmarks (Section 6.6).

2 Background

Next, we present background material necessary to understand the rest of the paper.

2.1 Knowledge Graphs

KGs represent structured knowledge as a directed graph $\mathcal{G}$ composed of a set of node entities $\mathcal{N}$ and their relational edges $\mathcal{E}$. Each edge encodes a fact that can be viewed as a triple (h, r, t), where h and t are the head and tail entities, and r is the relation linking them. For example, `(Paris, capital-of, France)` asserts that Paris is the capital of France. A path p in a KG is a sequence of connected triples that forms a relational chain between two entities. A length-N path is defined as:

$$p^N \equiv (h_0, r_1, h_1), (h_1, r_2, h_2), \ldots, (h_{N-1}, r_N, h_N) \tag{1}$$

Some prominent general-purpose KGs include DBpedia (Auer et al., 2007), Wikidata, and Google's Knowledge Graph (Singhal, 2012), which was designed to enhance search relevance through entity linking and ranking. Enriched with ontologies that represent semantic relationships between entities (nodes) and edges, KGs facilitate complex queries and reasoning. The multi-hop paths allow the KG to capture higher-order relations between h_0 and h_N. For instance, a length-3 (alternatively, a 3-hop) path could be $p^3 =$ `(Marie Curie, educated-at, Sorbonne), (Sorbonne, located-in, Paris), (Paris, capital-of, France)`. This chain captures the higher-order relation that Marie Curie was educated at an institution in the capital of France. Over time, KGs have evolved from general-purpose resources to specialized graphs in critical fields; biomedical graphs, such as UMLS (Bodenreider, 2004), SemMedDB (Kilicoglu et al., 2012), Hetionet (Himmelstein et al., 2017), underpin advanced applications in clinical informatics by unifying heterogeneous data into semantically rich networks.

2.2 Unified Medical Language System (UMLS) Knowledge Graph

In biomedicine, KGs have become indispensable for integrating disparate sources, including literature, ontologies, and clinical records to enable precision medicine and complex reasoning. At the heart of our data generation pipeline is a medical KG extracted from the comprehensive UMLS ontology. UMLS integrates multiple health and biomedical vocabularies into a unified framework by aligning synonymous terms under standardized Concept Unique Identifiers (CUIs) and linking them through a curated set of semantic relations. For instance, terms like “myocardial infarction,” “heart attack,” and “MI” are all mapped to the same CUI. Past pioneering work (Bodenreider, 2004) has constructed and filtered a KG from UMLS by representing each CUI as a node and the semantic relationships, such as “treats,” “causes,” or “is a subtype of,” as directed edges between nodes. Thereafter, researchers (Yasunaga et al., 2021) combined the disease sub-part of the UMLS KG with DrugBank (Wishart et al., 2018) to create an expansive drug-and-disease database. We traverse the paths of this constructed KG to generate a curriculum of medical relations ordered by path length. Simple 1-hop paths correspond to simple medical factoids such as Aspirin → may-treat → Myocardial Infarction. In contrast, more complex multi-hop paths can support clinically meaningful vignettes. For example, the path Diabetes Mellitus → predisposes to → Kidney Diseases → causes → Anemia captures the reasoning chain that diabetes may lead to kidney disease, which in turn can cause anemia due to impaired erythropoietin production.

2.3 The International Classification of Diseases (ICD)

Our work investigates the emergence of domain-specific expertise in fine-tuned LMs, specifically as it manifests in medical reasoning. However, medicine is a broad field, encompassing a complex topology of specialized subfields, each dedicated to the diagnosis and treatment of distinct categories of health conditions. To be truly useful, a superintelligent medical specialist must not only grasp the broad scope of medicine but, more importantly, demonstrate the ability to reason effectively within narrow, highly specialized domains. It is, therefore, necessary to evaluate and benchmark the capabilities of medical specialists across these specialized domains. The ICD (Organization, 1992), a globally recognized taxonomy for recording, reporting, and analyzing health conditions, provides a natural structure for this effort (Fig. 4 top). We leverage the ICD framework to design a benchmark called ICD-Bench (Section 4), which systematically evaluates domain-specific medical reasoning along distinct axes of ICD disease types.

3 Bottom-Up Curriculum Generation For Language Models

Overview: First, we use a KG as a scaffold for generating closed-ended tasks that are grounded in the structured entities of the KG. To this end, we traverse local paths of a KG to construct grounded tasks in the form of **(question, answer)** pairs. We then extract high-quality thinking traces for each QA pair from a reasoning language model grounded in the traversed path. During training, we fine-tune an LM on a curated curriculum of **(question, thinking trace, answer)** triplets generated by this pipeline. At inference time, we scale inference-time compute on the curriculum-tuned model.

3.1 The Task-Generation Pipeline

The ultimate goal of our task-generation pipeline is to curate reasoning tasks using KGs, guided by three core design principles:

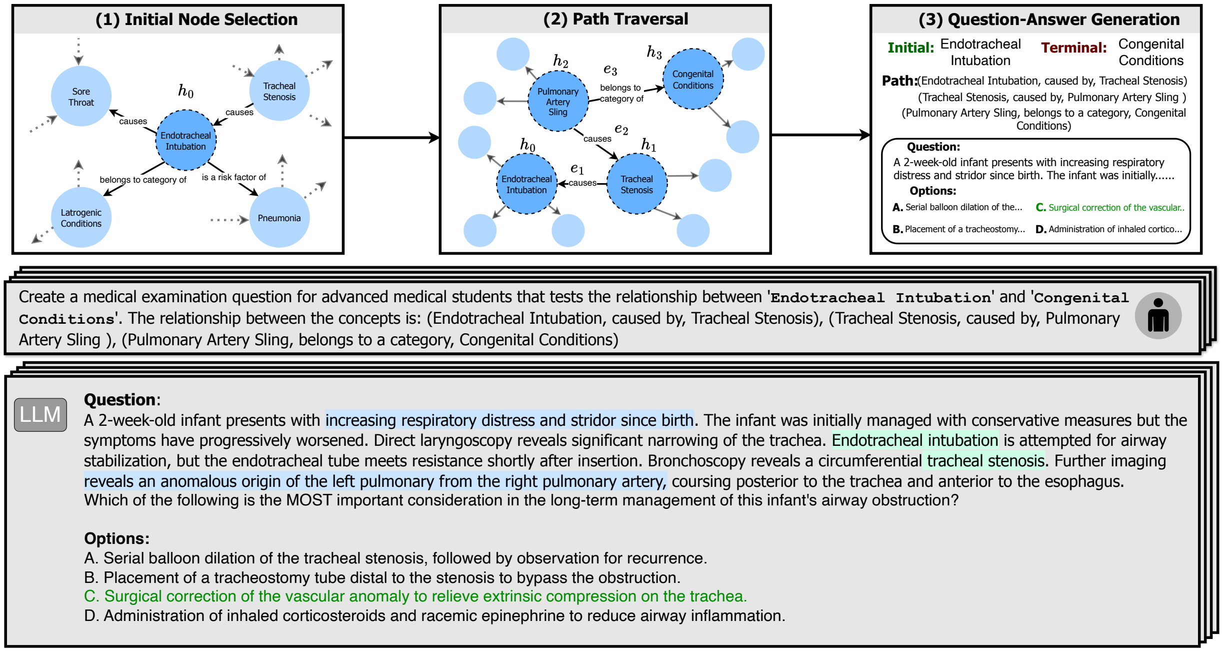
- **Closed-endedness**: Each generated task should have a distinct correct answer with the reasoning traceable to grounded paths in the KG.
- **Steerable Complexity**: The pipeline should facilitate reliable control over the depth of reasoning required to solve the generated tasks.
- **Diversity**: The pipeline should ensure that traced paths cover the entire KG instead of being concentrated on a few nodes.

3.1.1 Generating Grounded Question-Answering Tasks Using a KG

Each data point from our pipeline is instantiated as a multiple-choice question (MCQ) sourced from the KG. This QA format encourages reasoning as models must generate rationales for both identifying the correct answer and eliminating distractor options. The generation process, illustrated in Fig. 2 (top), comprises three main stages:

(1) Initial Node Selection: We begin by selecting an initial concept node h_0 from the KG.

(2) Path Traversal: From h_0, we sample a multi-hop path of length N on the KG in N steps. At each step t, we consider the set of all outgoing (relation, neighbor) pairs from the current node h_{t-1}, exclude any neighboring nodes already visited, and sample one pair (e_t, h_t) uniformly. This simultaneously selects the


(1) Initial Node Selection
Sore Throat
Tracheal Stenosis
h_0
causes
causes
Endotracheal Intubation
belongs to category of
is a risk factor of
Iatrogenic Conditions
Pneumonia
(2) Path Traversal
h_2
e_3
h_3
Pulmonary Artery Sling
belongs to category of
Congenital Conditions
e_2
causes
h_0
h_1
e_1
Endotracheal Intubation
causes
Tracheal Stenosis
(3) Question-Answer Generation
Initial: Endotracheal Intubation
Terminal: Congenital Conditions
Path:(Endotracheal Intubation, caused by, Tracheal Stenosis)
(Tracheal Stenosis, caused by, Pulmonary Artery Sling)
(Pulmonary Artery Sling, belongs to a category, Congenital Conditions)
Question:
A 2-week-old infant presents with increasing respiratory distress and stridor since birth. The infant was initially......
Options:
A. Serial balloon dilation of the...
C. Surgical correction of the vascular..
B. Placement of a tracheostomy...
D. Administration of inhaled cortico...
LLM

Create a medical examination question for advanced medical students that tests the relationship between '`Endotracheal Intubation`' and '`Congenital Conditions`'. The relationship between the concepts is: (Endotracheal Intubation, caused by, Tracheal Stenosis), (Tracheal Stenosis, caused by, Pulmonary Artery Sling), (Pulmonary Artery Sling, belongs to a category, Congenital Conditions)

Question:
A 2-week-old infant presents with increasing respiratory distress and stridor since birth. The infant was initially managed with conservative measures but the symptoms have progressively worsened. Direct laryngoscopy reveals significant narrowing of the trachea. Endotracheal intubation is attempted for airway stabilization, but the endotracheal tube meets resistance shortly after insertion. Bronchoscopy reveals a circumferential tracheal stenosis. Further imaging reveals an anomalous origin of the left pulmonary from the right pulmonary artery, coursing posterior to the trachea and anterior to the esophagus.
Which of the following is the MOST important consideration in the long-term management of this infant's airway obstruction?

Options:
A. Serial balloon dilation of the tracheal stenosis, followed by observation for recurrence.
B. Placement of a tracheostomy tube distal to the stenosis to bypass the obstruction.
C. Surgical correction of the vascular anomaly to relieve extrinsic compression on the trachea.
D. Administration of inhaled corticosteroids and racemic epinephrine to reduce airway inflammation.

Figure 2: **Generating QA tasks from a KG path.** We explore KG paths to derive QA pairs grounded in the KG. We choose a KG path by starting from an initial node (left) and iteratively sampling (relation, entity) pairs from the current node's neighbors to obtain an N-hop path p^N (middle). The sampled path is mapped to a natural language QA task by prompting a backend LLM (right). The bottom of the diagram shows an example of a generated QA pair, where text highlighted in green indicates entities revealed in the question and blue highlights indicate latent entities. Effectively solving the QA task requires recalling latent entities and reasoning along the KG path to reach the correct solution.

relation e_t and the next node h_t in one draw. More formally, given h_0 and path-length N:

```
Initialize Path: $p^0 = \emptyset$
For $t = 1$ to $N$ :
    Gather Candidates: $\mathcal{C}_t = \{(e, v) \mid (h_{t-1}, e, v) \in \text{Neighbors}(h_{t-1}),\ v \notin \{h_0, \dots, h_{t-1}\}\}$
    Sample Next Hop: $(e_t, h_t) \sim \text{Uniform}(\mathcal{C}_t)$
    Extend Path: $p^{t+1} \leftarrow p^t \cup (h_{t-1}, e_t, h_t)$
```

Here, $\mathcal{C}_t$ collects all valid outgoing pairs (e, v) from the current node h_{t-1}, excluding those whose target v has already appeared in $\{h_0, \dots, h_{t-1}\}$. After N hops, the complete path is $(h_0, r_1, h_1), \cdots, (h_{N-1}, r_N, h_N)$.

(3) Question-Answer Generation: Our pipeline transforms each sampled KG path into a question-answering task by leveraging a backend LLM. Specifically, we design a template prompt that tasks the model with constructing a vignette (a clinical one in the case of medical superintelligence) and posing an MCQ whose resolution depends on traversing the entire path. The template

1. instructs the model to formulate a question that links the initial node h_0 to the terminal node h_N,
2. provides the complete path as context to ensure factual grounding, and
3. enforces a correct answer with other plausible but false options.

The transformation can be formalized as follows:

$$\text{Question, Answer} = LM\big(\,\mathcal{T}\big(\text{ source} = h_0,\ \text{target} = h_N\ , \text{context} = p^N\big)\big)$$

Here, $\mathcal{T}(.)$ denotes the template-based prompt (see Appendix A.2) and LM(.) denotes sampling from an LLM. We use the `Gemini 2.0 Flash` model (Google DeepMind, 2025b) to generate the QA pairs. Fig. 2 (bottom) shows a QA example along with its path context. Returning to the aforementioned design principles, the QA format naturally endows closed-endedness since the model must provide a single correct answer, the source node and the subsequently sampled path provide a natural control over the diversity of the QA task, while the path length allows for steering complexity.

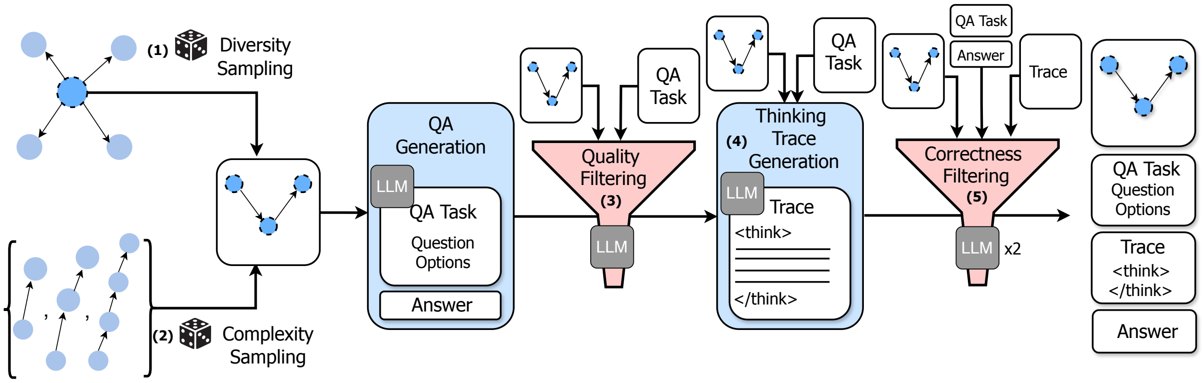


Figure 3: **Overview of our curriculum curation pipeline.** Starting from KG-derived paths, we sample for node diversity and path complexity (Steps 1-2), followed by quality filtering of generated QA pairs (Step 3). We then generate thinking traces using a strong reasoning LLM grounded in the KG path (Step 4), and finally perform correctness filtering using two independent grader LLMs to ensure factual trace alignment to the KG path and answer validity (Step 5).

3.1.2 Curriculum Curation

We use the transformation method to leverage the KG path to QA pair to assemble a training curriculum of high-quality QA pairs. The dataset is carefully curated for *diversity* of source nodes across the KG, *complexity* via enforcing different-length hops, and *quality and correctness* by introducing filtering heuristics for QA coherence. The pipeline, illustrated in Fig. 3, proceeds via the following steps.

(1) Diversity Sampling: To ensure that our sampled paths provide broad coverage of the KG and avoid clustering around a small subset of highly-connected nodes, we enforce diversity while selecting the source node. During the data curation process, we maintain a running selection frequency with which nodes are sampled on paths. Then we sample the source node based on the inverse of the selection frequency, ensuring that unsampled nodes or fewer-sampled nodes are sampled more. Let f_i denote the sample frequency of node i in the set of generated QA pairs. Then node i is sampled as the source node with probability:

$$p_i = \frac{w_i}{Z}\ ; \quad \text{where } w_i = \frac{1}{f_i + \epsilon},\ Z = \sum_{i=1}^{|\mathcal{N}|} w_i$$

ϵ (= 1) is a small constant that prevents division by zero for unsampled nodes.

(2) Complexity Sampling: To induce a graded notion of reasoning difficulty, we uniformly sample KG paths of varying lengths from $\{1, \cdots, N\}$ instead of always selecting the longest possible paths. This introduces a natural curriculum where shorter paths typically yield recall-based or single-hop queries, while longer paths require multi-hop, compositional reasoning. By training on a range of path complexities, the model develops balanced reasoning skills and avoids overfitting to long, potentially noisy chains, thereby improving robustness and generalization. We ablate and show the effect of sampled KG depth on reasoning performance in Section 6.4.

(3) Quality Filtering: To ensure high-quality QA pairs, we implement a multi-stage filtering process. We first discard generations with application programming interface (API) call failures, incomplete responses, or distracting artifacts, such as ASCII strings or code blocks. Next, we enforce strict adherence to a predefined QA template. Each question must be phrased as a vignette grounded in a KG path, followed by one correct answer and three plausible distractors, with consistent formatting (e.g., option labels like 'A.'). Finally, we eliminate QA pairs with low-quality distractors, such as near-duplicates or distractors that closely resemble the correct answer, to preserve the discriminative integrity of each question.

(4) Thinking Trace Generation: After QA pairs pass quality filters, we distill high-quality thinking traces in natural language from the underlying KG paths. For each retained question, we prompt a strong reasoning model with the vignette and options from the QA pair with the full KG path as context, as follows:

$$\text{Thinking Trace} = LM\big(\text{question} = (\text{Vignette, Options}), \text{context} = p^N\big)$$

More specifically, we distill the traces from the `Gemini-2.5-Pro` model (Google DeepMind, 2025c), which has demonstrated state-of-the-art reasoning capabilities. We instruct the model to reason through the question to infer the correct answer, referencing the KG path (see Appendix A.3, Prompt 2). By anchoring the reasoning trace to the KG path, we produce structured rationales grounded in the KG that ensure strong relational supervision for fine-tuning.

(5) Correctness Filtering: Despite grounding questions in explicit KG paths, errors can arise due to ambiguous phrasing, inconclusive evidence along the KG path, or LLM hallucinations. To address this, we perform a final correctness check to ensure that each QA item (question, thinking trace, answer) is unambiguously interpretable based on the provided path and that the thinking trace faithfully follows the path to arrive at the correct answer without hallucinations. Since the KG contains relations curated by experts, ensuring the logical grounding of the QA item with its explanation in the KG path also ensures the scientific factualness of the QA item and the explanation. We organize the complete context of the QA item under a template prompt and task an LLM grader to verify correctness. We specifically instruct the grader model (see Appendix A.3, Prompt 3) to evaluate whether (a) the correct answer follows from the vignette and the KG path, and (b) every claim in the thinking trace is supported by the KG path, without hallucinations. The grader outputs a binary verdict. To guard against idiosyncratic failures of any single model, we enforce a two-factor agreement using two grader models – `Gemini 2.0 Flash` and `Qwen 2.5-72B` (Yang et al., 2025b). We retain QA items only if both independent grader models verify correctness, ensuring robustness through cross-model consistency.

We iteratively repeat these steps until we have a user-defined size of high-quality QA items. The task-generation pipeline has been summarized in Appendix A.3 Algorithm 1.

3.2 Curriculum Tuning and Inference

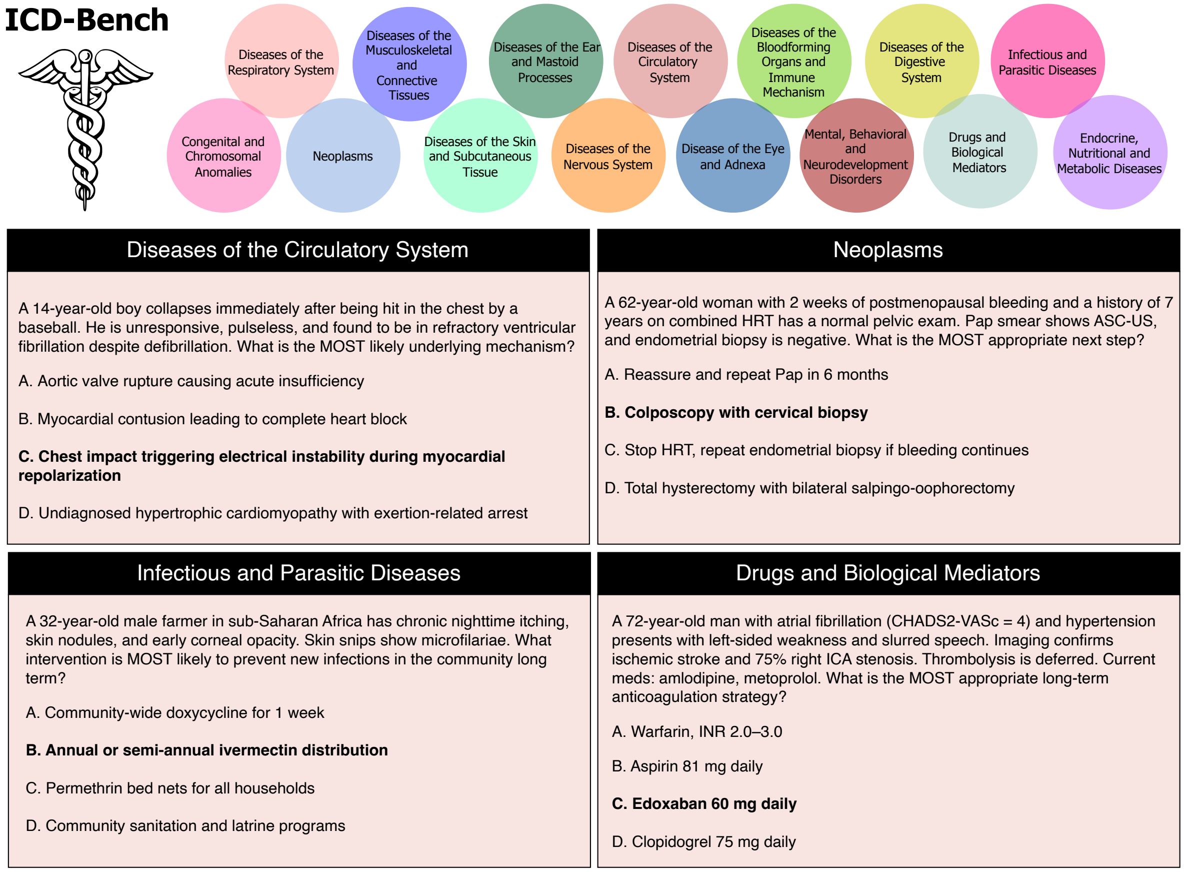
Curriculum Tuning: We use our curated dataset, specifically reasoning traces derived from KG paths, to perform supervised fine-tuning (SFT) of off-the-shelf LMs via the next token prediction objective. Prior to training, we map each (question, thinking trace, answer) datapoint to a chat template, with the thinking trace inserted between special `<think>` and `</think>` delimiters to signal the beginning and end of the thinking process, respectively. We refer to the resulting fine-tuned models as *curriculum-tuned* models, since they have been explicitly trained to acquire a structured reasoning curriculum grounded in KG primitives.

Inference: At inference time, we scale compute on our curriculum-tuned models by expanding the generated thinking trace (Snell et al., 2024), either by generating multiple traces in parallel (Wei et al., 2022; Wang et al., 2023) or by extending individual traces through iterative refinement (Muennighoff et al., 2025). Specifically:

- **Parallel Scaling:** We generate n independent thinking traces for each test question in parallel. Each instance produces a complete trace (delimited by `<think>` and `</think>`), followed by an answer. The final prediction is obtained via majority voting across the n outputs.
- **Iterative Refinement:** We also allocate additional compute to trace refinement, encouraging the model to re-evaluate its reasoning. Following prior work (Muennighoff et al., 2025), we intervene in the decoding process by replacing the end-of-thinking delimiter `</think>` with prompts like `hmm, let's double check`, prompting the model to continue its thought process before finalizing an answer.

4 ICD-Bench Evaluation Suite

To rigorously evaluate the emergence of domain-specific reasoning capabilities in our curriculum-tuned models, we introduce **ICD-Bench**, a targeted QA benchmark aligned with the ICD taxonomy (see Section 2.3). The benchmark is constructed using structured knowledge embedded in the UMLS KG and is designed to evaluate models on domain-specific tasks spanning medical sub-specialties. We begin by describing its construction.


ICD-Bench
Diseases of the Respiratory System
Diseases of the Musculoskeletal and Connective Tissues
Diseases of the Ear and Mastoid Processes
Diseases of the Circulatory System
Diseases of the Bloodforming Organs and Immune Mechanism
Diseases of the Digestive System
Infectious and Parasitic Diseases
Congenital and Chromosomal Anomalies
Neoplasms
Diseases of the Skin and Subcutaneous Tissue
Diseases of the Nervous System
Disease of the Eye and Adnexa
Mental, Behavioral and Neurodevelopment Disorders
Drugs and Biological Mediators
Endocrine, Nutritional and Metabolic Diseases

Diseases of the Circulatory System

A 14-year-old boy collapses immediately after being hit in the chest by a baseball. He is unresponsive, pulseless, and found to be in refractory ventricular fibrillation despite defibrillation. What is the MOST likely underlying mechanism?

A. Aortic valve rupture causing acute insufficiency

B. Myocardial contusion leading to complete heart block

C. Chest impact triggering electrical instability during myocardial repolarization

D. Undiagnosed hypertrophic cardiomyopathy with exertion-related arrest

Neoplasms

A 62-year-old woman with 2 weeks of postmenopausal bleeding and a history of 7 years on combined HRT has a normal pelvic exam. Pap smear shows ASC-US, and endometrial biopsy is negative. What is the MOST appropriate next step?

A. Reassure and repeat Pap in 6 months

B. Colposcopy with cervical biopsy

C. Stop HRT, repeat endometrial biopsy if bleeding continues

D. Total hysterectomy with bilateral salpingo-oophorectomy

Infectious and Parasitic Diseases

A 32-year-old male farmer in sub-Saharan Africa has chronic nighttime itching, skin nodules, and early corneal opacity. Skin snips show microfilariae. What intervention is MOST likely to prevent new infections in the community long term?

A. Community-wide doxycycline for 1 week

B. Annual or semi-annual ivermectin distribution

C. Permethrin bed nets for all households

D. Community sanitation and latrine programs

Drugs and Biological Mediators

A 72-year-old man with atrial fibrillation (CHADS2-VASc = 4) and hypertension presents with left-sided weakness and slurred speech. Imaging confirms ischemic stroke and 75% right ICA stenosis. Thrombolysis is deferred. Current meds: amlodipine, metoprolol. What is the MOST appropriate long-term anticoagulation strategy?

A. Warfarin, INR 2.0–3.0

B. Aspirin 81 mg daily

C. Edoxaban 60 mg daily

D. Clopidogrel 75 mg daily

Figure 4: **ICD-Bench evaluation suite.** Top: The 15 medical sub-specialties derived from the ICD-10 taxonomy, each corresponding to a distinct category in the benchmark. Each node in the UMLS KG is mapped to one or more of these categories to guide domain-specific QA generation. Bottom: Sample QA items drawn from different ICD categories, illustrating the diversity of the benchmark in medical reasoning tasks, from treatment selection and diagnostic evaluation to mechanistic and public health interventions across disease types.

4.1 ICD-Bench Construction Procedure

(1) Aligning the KG to the ICD-10 Taxonomy: To enable fine-grained control over domain-specific QA generation, we map each node in the UMLS KG to one or more of the 15 ICD categories illustrated in Fig. 4 (top). This mapping is performed by an LLM classifier, which assigns categories only to nodes with a strong affinity.

(2) Question Generation: We use the stratified KG to generate hop-controlled QA items per category, using our QA generation method (Section 3.1.1), as follows:

1. **Select Category:** Choose category C from the ICD-10 taxonomy.

2. **Sample Source Node:** Sample a source node h^0 belonging to category C.
3. **Select Path Complexity:** Choose path complexity N.
4. **Generate QA:** Traverse a length-N KG path p^N beginning at h^0 and generate a QA pair.

(3) Quality and Correctness Filtering: Finally, we subject the generated QA pairs to quality and correctness checks using the steps outlined in Section 3.1.2.

4.2 ICD-Bench Composition

The final ICD-Bench evaluation suite comprises 3,675 high-quality QA items, systematically generated through the controlled pipeline described above. These items are evenly distributed across the 15 ICD-10 categories, enabling a balanced assessment of reasoning capabilities across diverse medical domains. Each category contributes 245 QA items, stratified by 100 questions derived from two-hop KG paths, 100 from three-hop paths, 30 from four-hop paths, and 15 from five-hop paths. Tasks from one-hop paths are omitted to minimize bias towards those that require simple recall or memorization, rather than deeper reasoning. This structure ensures that ICD-Bench probes both the breadth of domain-specific knowledge and the depth of compositional reasoning needed to navigate each domain. Fig. 4 (bottom) shows representative QA examples drawn from distinct ICD categories, spanning simple entity-relational queries and composite chains involving treatments, diagnostics, and etiological factors.

5 Experiment Setup

Next, we define the experimental setup.

5.1 Training Curriculum Curation and Decontamination Setup

Curriculum Curation: We leverage our proposed task-generation pipeline (Section 3.1) to curate a training curriculum of 24,000 QA tasks along with their thinking traces on the UMLS KG. The generated tasks span diverse medical entities and relations on the KG, and are distributed uniformly over multi-hop lengths $\in \{1,\ 2,\ 3\}$. We restrict hop lengths to $N \leq 3$ based on an empirical observation that paths longer than three hops often traverse semantically unmeaningful and weak relations, diminishing the coherence and correctness of the resulting questions. Moreover, we reserve the small subset of semantically meaningful long-range paths ($N \geq 4$) for evaluation via ICD-Bench.

Decontamination: We also perform a two-fold decontamination of our generated tasks prior to their inclusion in the training data. First, we exclude any QA pair where the underlying KG path *exactly* traverses a KG path of any ICD-Bench QA pair. This prevents memorized KG paths from contaminating evaluations. However, we allow *partial* path overlaps, as our goal is to allow models to learn and generalize from individual KG primitives. Therefore, we additionally adopt an 18-gram threshold for the text overlap filter to eliminate QA pairs that are highly similar to those in ICD-Bench, filtering out close matches while preserving distinct yet conceptually related questions.

Fig. 5 displays the breakdown of category entities (left) and relations (right) in the KG curriculum, and the distribution of the thinking trace lengths across multi-hop paths (bottom). Examples of generated tasks and token breakdown of the curriculum are provided in Appendix B.

5.2 Curriculum-Tuning Setup

Our work investigates how domain-specific reasoning abilities can emerge in a general-purpose model through fine-tuning on structured curricula derived from a domain KG. We adopt the open-source QwQ-32B LM (Team, 2025b) as our base, leveraging its strong reasoning foundations acquired via large-scale reinforcement learning. To study how a curriculum affects generalization, we construct three training datasets that progressively incorporate tasks from deeper KG paths and fine-tune the base model on each dataset under a fixed floating-point operations (FLOPs) budget. This yields three fine-tuned models:

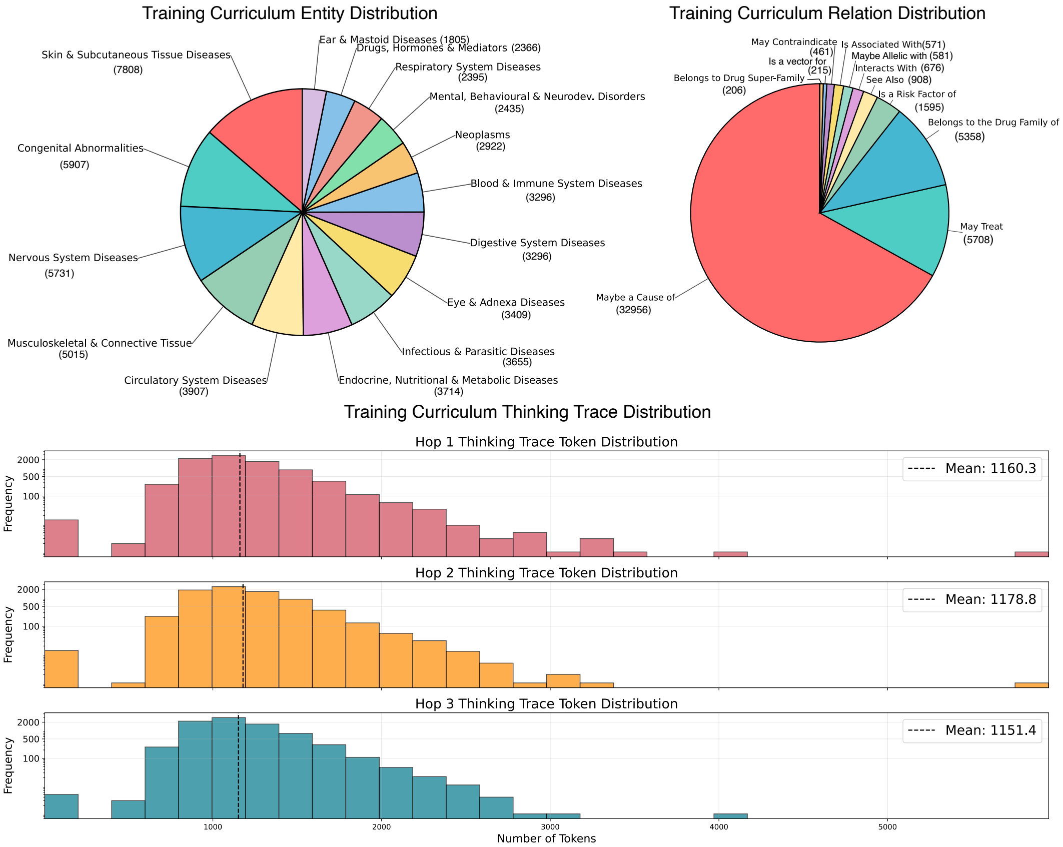


Figure 5: **Distributional statistics of the curated training curriculum.** The dataset spans 24,000 QA items grounded in UMLS KG paths. Left: Breakdown of entities in the sampled KG paths, across the ICD categories they belong to. Right: Distribution of relation types along sampled KG paths, spanning different semantic relations. Bottom: Hop-wise distribution of thinking trace lengths across the dataset, reflecting variance in reasoning complexity.

- **QwQ-Med-1:** Fine-tuned on 8,000 tasks derived from single-hop KG paths for 24 epochs.
- **QwQ-Med-2:** Trained on 16,000 tasks that combine one-hop and two-hop paths for 12 epochs.
- **QwQ-Med-3:** Extended to include three-hop paths, totaling 24,000 tasks over 8 epochs.

From a pedagogical standpoint, each successive model is trained on a curriculum that grows both deeper and wider with respect to the underlying KG. Depth increases through the inclusion of longer multi-hop reasoning chains, while breadth expands as the model is exposed to a more diverse set of KG entities and relational contexts. All models are fine-tuned using low rank adapters (LoRA) (Hu et al., 2021) with rank 16 on 8 H100 NVIDIA GPUs, with each run taking approximately 20 hours. The complete SFT-through-LoRA setup is outlined in Appendix C.

6 Experiments

Next, we present our experiments.

6.1 Understanding Inference-Time Scaling Behavior of Curriculum-Tuned Models

(S1) Setup: We evaluate our curriculum-tuned models on ICD-Bench by extending inference-time compute using the parallel and iterative refinement strategies outlined in Section 3.2. For parallel inference, we vary

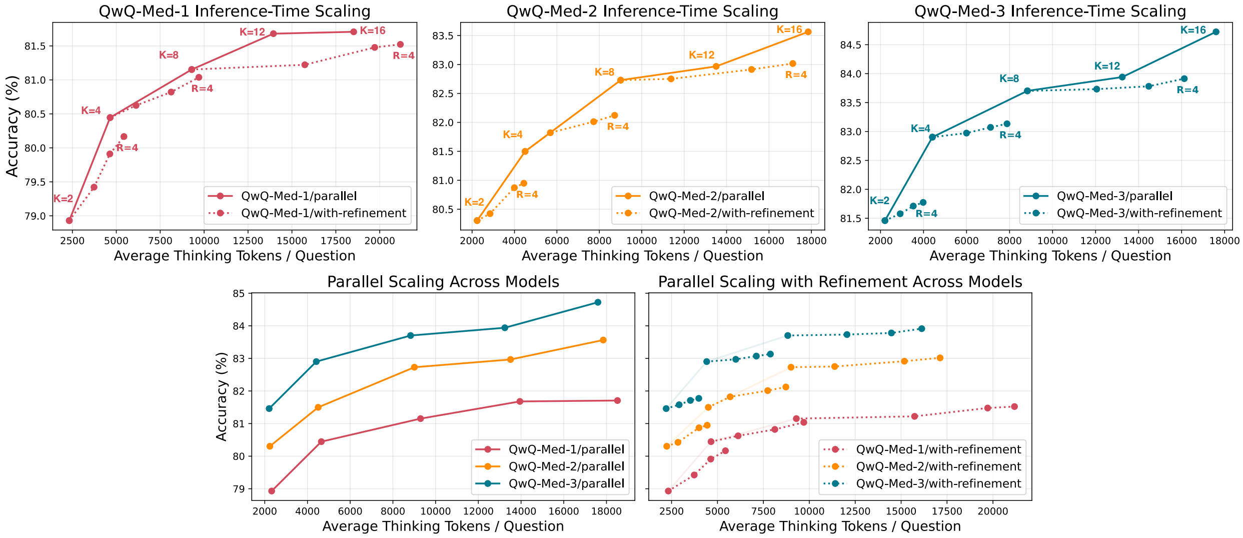


Figure 6: **Inference-time scaling curves for curriculum-tuned models on ICD-Bench.** Top row: Each plot displays per-model curves where solid lines denote pure parallel scaling and dotted lines denote iterative refinement augmentation. Deeper curriculum models (QwQ-Med-2, QwQ-Med-3) benefit more from parallel scaling, while QwQ-Med-1 remains amenable to refinement. Bottom row: Each plot shows a comparison of per-scaling technique curves. QwQ-Med-3 trained on the entire generated curriculum demonstrates compute-optimality over models trained on partial curricula. Bootstrapped confidence intervals over 500 samples were $< 0.75\%$.

the number of concurrent reasoning streams with $K \in \{2, 4, 8, 12, 16\}$, setting decoding temperature to 0.6. For iterative refinement, we further augment each parallel stream with $R = 4$ refinement steps, evaluating this setting for $K \in \{2, 4, 8\}$. We evaluate each model-compute configuration on the full set of 3,675 ICD-Bench questions across 15 medical categories. We report the overall accuracy alongside the average number of thinking tokens consumed per question. The inference-time scaling curves for our models are shown in Fig. 6. We observe that:

(O1.1) Parallel scaling outperforms iterative refinement with increasing curriculum. QwQ-Med-2 and QwQ-Med-3 exhibit steeper gains from parallel scaling (solid lines) while refinement (dotted lines) saturates, unlike QwQ-Med-1, where improvements from parallelism asymptote and converge with refinement (top row, Fig. 6). This dichotomy indicates that deeper curriculum models, having acquired structured and diverse KG primitives, benefit more from exploring multiple reasoning paths in parallel. In addition, unlike math and coding tasks, where refinement helps due to verifiable intermediate steps, medical diagnosis hinges on early differential diagnosis, making parallel sampling more impactful for curriculum-tuned models.

(O1.2) Structured curriculum-tuning enables inference-time compute optimality. With an increasingly difficult curriculum, our fine-tuned models achieve higher accuracy at lower inference budget, as reflected by leftward shifts in the scaling curves (bottom row, Fig. 6). Despite being fine-tuned on an equal training FLOPs budget, models learn to allocate inference budget more effectively. We posit that exposure to deeper multi-hop chains and broader KG coverage allows models to acquire and reuse reasoning derived from KG primitives, which enables them to converge to accurate answers with less iterative search or brute-force sampling at test time.

In the rest of the paper, we use parallel inference-time scaling unless otherwise mentioned.

6.2 Domain-Specific Reasoning Emerges from Curriculum-Tuned Models

(S2) Setup: We gain insight into the domain-specific capabilities acquired through curriculum tuning by evaluating our models on the category-specific branches of ICD-Bench. Each subset comprises 245 medical

QA examples focused on a specific disease type, allowing us to disentangle performance gains along distinct clinical axes.

Baselines: We compare our curriculum-tuned models against four baselines: (1) the QwQ-32B base model, which serves as our reference general-purpose reasoning model, (2) DeepSeek-R1-Distilled Qwen (DeepSeekAI, 2025), another strong open-source reasoning model distilled from Deepseek-R1, and (3) two proprietary state-of-the-art reasoning models, o3 (OpenAI, 2025c) and Gemini-2.5-Pro (Google DeepMind, 2025c), known for strong domain generalization and competitive benchmark performance. We perform inference-time scaling on the open-source models while reporting pass@1 accuracy (of the first generated solution) for the proprietary models.

In Fig. 1, we present the performance of our model relative to proprietary baselines across ICD-Bench categories. Fig. 7 shows scaling results of reasoning models across ICD-Bench categories. Our major takeaways are:

(O2.1) Curriculum-tuned models significantly outperform other reasoning models. Our models consistently outperform all open-source baselines across inference budgets by 10-20%. Moreover, open-source reasoning models tend to plateau early with increasing compute, whereas our curriculum-tuned models demonstrate a better utilization of inference-time budget. Strikingly, our models also outperform o3 and Gemini-2.5-Pro, despite their massive model size and training on web-scale data. While proprietary models are competitive in highly prevalent disease cases, like neoplasms, circulatory, and respiratory conditions, that are more frequently represented in text corpora, our curriculum-tuned models show crucial improvements in less prevalent categories, like congenital abnormalities and nervous system disorders. Finally, the hop-1 curriculum training itself confers a substantial advantage by endowing curriculum-tuned models with reliable composable primitives, and the subsequent curriculum enables the tackling of more difficult problems, explored more comprehensively in Section 6.3.

(O2.2) Expanding the curriculum improves domain-specific reasoning. Across most ICD-Bench categories, QwQ-Med-3 outperforms both QwQ-Med-2 and QwQ-Med-1, highlighting the cumulative benefit of scaling curriculum depth and diversity. In some categories, QwQ-Med-2 and QwQ-Med-3 perform similarly, which may reflect early saturation from acquiring densely linked support KG paths that already capture the central reasoning primitives needed for that disease type.

Qualitative. In Examples 1 and 2, we present sample responses from our QwQ-Med-3 model. Each entity and relation from the underlying KG path is color-coded distinctly. Corresponding segments in the model's response that recall these entities or trace the reasoning along the relations are highlighted in the same color. This alignment illustrates that curriculum-tuned models can recall acquired KG primitives and coherently compose them during reasoning. Additional outputs and comparison to the base model are provided in Appendix D.

6.3 Curriculum-Tuned Models Improve Robustness to Task Difficulty

Real-world clinical reasoning tasks, such as those in ICD-Bench, exhibit a spectrum of complexity, from direct factual recall to implicit, multi-step inference. In this section, we probe whether our curriculum-tuned models exhibit improvements across this full difficulty range, specifically whether the acquisition of KG primitives enables them to reliably reason on hard tasks.

Task Difficulty Estimation: To estimate difficulty, we use the base QwQ model as a proxy evaluator. For each task, we compute its pass@1 rate, the fraction of times the model produces a correct answer across 16 independently sampled generations. This score serves as a difficulty heuristic, with lower rates signifying harder tasks. The resulting distribution is bimodal, with a dominant mass near 100% and a secondary mode at lower success rates. Subsequently, we partition the tasks into five difficulty bins based on this empirical distribution, capturing a fine-grained spectrum of task hardness (see Appendix E for difficulty distribution and difficulty bin cutoffs).

(S3) Setup: We stratify our ICD-Bench evaluation across difficulty bins and report in Fig. 8 the accuracy of curriculum-tuned models and previous baselines under the full inference budget setting. We find:

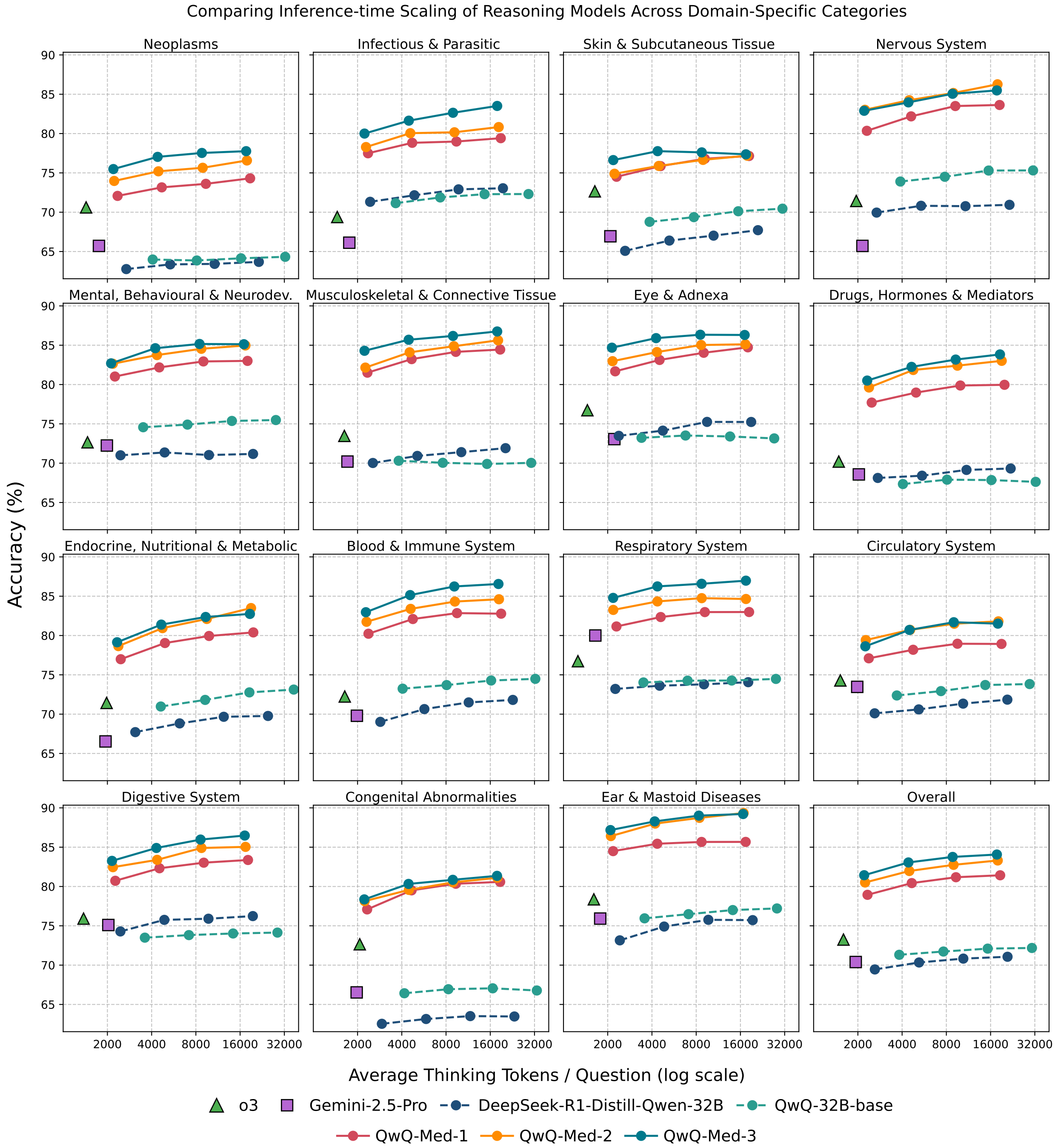


Figure 7: **Domain-specific performance of our curriculum-tuned models across ICD-Bench categories.** Curriculum-tuned models significantly outperform proprietary and open-source reasoning baselines across medical sub-specialities. Moreover, curriculum-tuned models demonstrate improved inference-time scaling, with greater curriculum (QwQ-Med-3) enabling compute-optimality across categories. Bootstrapped confidence intervals over 500 samples were $< 0.75\%$.

(O3.1) Task difficulty is symmetrical across models. All models exhibit a consistent drop in accuracy as task difficulty increases (Fig. 8, left), validating our estimator. While open-source models like R1-Distilled Qwen and the base QwQ perform well on easy tasks, their performance significantly degrades on harder ones. Curriculum-tuned models, by contrast, outperform baselines across all bins, with the performance gap with

Example 1: QwQ-Med-3 Response

KG Path from Diseases of the Respiratory System:

Pulmonary alveolar microlithiasis $\xrightarrow{\text{maybe causes}}$ Pulmonary hypertension $\xrightarrow{\text{treated by}}$ Nitric oxide

Question:

A 35-year-old male presents with progressive dyspnea, a loud P2, right ventricular heave, and sand-like pulmonary opacities. CT shows diffuse microliths. Echo reveals RV hypertrophy and pulmonary artery systolic pressure of 65 mmHg. Which of the following is the MOST appropriate therapy?

Options:

A. Inhaled corticosteroids B. **Sildenafil** C. N-acetylcysteine D. Bronchodilators

QwQ-Med-3 Response:

`<think>` Okay, let's break down this case step-by-step to figure out the best treatment. **1. Deconstruct the Presentation:** .... point clearly towards right heart strain, specifically right ventricular failure due to pulmonary hypertension .... **2. Interpret the imaging:** The chest X-ray shows those classic "sand-like" opacities. This is a huge clue! It almost always makes us think of Pulmonary Alveolar Microlithiasis (PAM) .... **3. Connect the dots:** So, we have PAM confirmed. These little stones build up in the alveoli, basically filling them up .... Chronic lung disease like PAM is a well-known cause of Group 3 Pulmonary Hypertension, also called hypoxic pulmonary hypertension .... Echocardiography shows right ventricular hypertrophy and an estimated pulmonary artery systolic pressure of 65 mmHg. .... **6. Evaluate the options:** .... B. Sildenafil: This is a Phosphodiesterase-5 (PDE5) inhibitor. PDE5 inhibitors are specifically used to treat pulmonary hypertension by causing vasodilation (relaxing) of the pulmonary arteries. .... standard therapy for PH, including Group 3 PH secondary to lung diseases like PAM. `</think>` Final Answer: **B**

the base widening at higher difficulties. This divergence suggests that our models are going beyond factual recall to internalizing structured reasoning capabilities that generalize to challenging scenarios while the base models exhibit surface-level knowledge but struggle to compose their reasoning to solve complex, unseen tasks.

(O3.2) Curriculum-tuned models show greater marginal gains on harder tasks. We further analyze the marginal contributions of successive curriculum-tuned models over the base (Fig. 8, right). On easy tasks, most of the performance is already captured by the base model, indicating that KG primitives are inessential for most easy tasks. However, starting with medium-difficulty tasks, the base model contributes less than half the final performance, with QwQ-Med-1 bridging much of the remaining gap. Crucially, for the hardest tasks, the base model offers near-zero accuracy, with the final gains on these tasks coming almost entirely from QwQ-Med-3, indicating that exposure to deeper, compositional paths in the KG is necessary for bridging the gap between recall and reasoning, and for enabling models to tackle hard domain-specific reasoning tasks.

6.4 Disentangling the Effect of Curriculum Depth from Curriculum Diversity

The effectiveness of a curriculum-tuned model stems from training on reasoning data derived by exploring (1) multi-hop KG paths of different lengths over (2) diverse entities of the KG. We determine the contribution of each factor to the overall performance by incrementally ablating over the data mixture and size used for curriculum tuning.

(S4) Setup: We begin with an 8,000-sample dataset derived solely from single-hop KG paths, representing a shallow but diverse baseline curriculum. To assess the effect of depth, we construct a second 8,000-sample dataset consisting entirely of three-hop paths, holding diversity constant. A third variant samples uniformly from multi-hop paths of length $1, 2, 3$ using our proposed complexity-sampling procedure. Finally, we scale to 24,000 samples via additional diversity sampling to maximize KG coverage. We fine-tune a model on each dataset under the same training FLOPs budget, and show their evaluation results on ICD-Bench in Fig. 9. Our major findings are:

Example 2: QwQ-Med-3 Response

KG Path from Diseases of the Nervous System:

Ethinylestradiol $\xrightarrow{\text{is a risk factor of}}$ Cerebrovascular accident $\xrightarrow{\text{maybe causes}}$ Gliosis

Question:

A 38-year-old woman with a history of migraines presents to her neurologist for follow-up. She has been taking an oral contraceptive containing ethinylestradiol for the past 10 years. ... MRI shows multiple small areas of T2 hyperintensity ... no evidence of hypertension or other cardiovascular risk factors. What is the MOST likely pathological process?

Options:

A. Demyelination B. Apoptosis C. **Reactive gliosis** D. Amyloid angiopathy

QwQ-Med-3 Response:

`<think>` Okay, let's break down this question step-by-step .... **3. Consider the Role of Ethinylestradiol**. The patient has been on Ethinylestradiol for a long time. Ethinylestradiol is an estrogen. We need to think about the potential effects of long-term estrogen exposure .... **4. Connect Estrogen and Migraines:** It's well-known that estrogen can have complex effects on migraines .... **5. Link Migraines and White Matter Changes:** .... recognized association between chronic migraines (especially migraine with aura) and the presence of these kinds of white matter lesions. The exact mechanism isn't fully understood .... **6. Evaluate the Options based on the Pathophysiology:** .... Reactive gliosis resulting from subclinical ischemic events: The white matter T2 hyperintensities represent areas of injury. This injury could be related to the migraine mechanism itself .... The body's response to this injury involves glial cells proliferating and changing their properties .... this is called gliosis. .... **7. Conclude:** The most plausible explanation links the patient's chronic migraines (potentially exacerbated by long-term ethinylestradiol use) to subclinical vascular or ischemic events in the brain. These events lead to tissue injury and the subsequent reactive gliosis, which manifests as T2 hyperintensities on MRI. `</think>` Final Answer: **C**

(O4.1) Our full task-curation pipeline is important. As shown in Fig. 9 (left), each stage of our task-curation pipeline is crucial to overall performance. Incorporating deeper KG paths alone yields a substantial gain, while holding diversity constant, evident in the improvement from the single-hop to the three-hop dataset. However, depth alone is not sufficient: sampling a balanced mix of paths via our complexity sampling yields a further improvement, suggesting that exclusive exposure to long paths may lead to overfitting. Lastly, scaling the dataset to 24,000 samples through additional diversity sampling also yields improvement, though its impact is less pronounced than path depth.

(O4.2) Compute-optimal KG depth depends on task difficulty. In Fig. 9 (right), we break down model performance on ICD-Bench by task difficulty, as estimated with the method presented in Section 6.3, comparing the relative gains of KG depth, complexity sampling, and diversity sampling over a single-hop baseline. Diversity sampling consistently improves performance across all difficulty levels. However, the optimal KG depth varies significantly with task difficulty. On easy tasks, using only three-hop chains slightly degrades performance, whereas the balanced dataset offers a modest improvement. For medium-difficulty tasks, a balanced mix of paths achieves near-optimal performance over maximizing path length. In contrast, on the hardest tasks, the three-hop-only dataset outperforms the balanced dataset. Together, these trends suggest that when task difficulty is known *a priori*, the compute-optimal curricula depth should be composed accordingly: shallow paths suffice for easier tasks, moderate complexity benefits intermediate reasoning, and deep multi-hop traces become critical only at the hardest levels.

6.5 Curriculum-Tuned Models Bridge the Recall-Reasoning Gap

Domain-specific reasoning hinges on being able to reliably recall relevant entities and relations, and then reason over them to reach a correct conclusion. We gain insight into the ability of our models to bridge recall

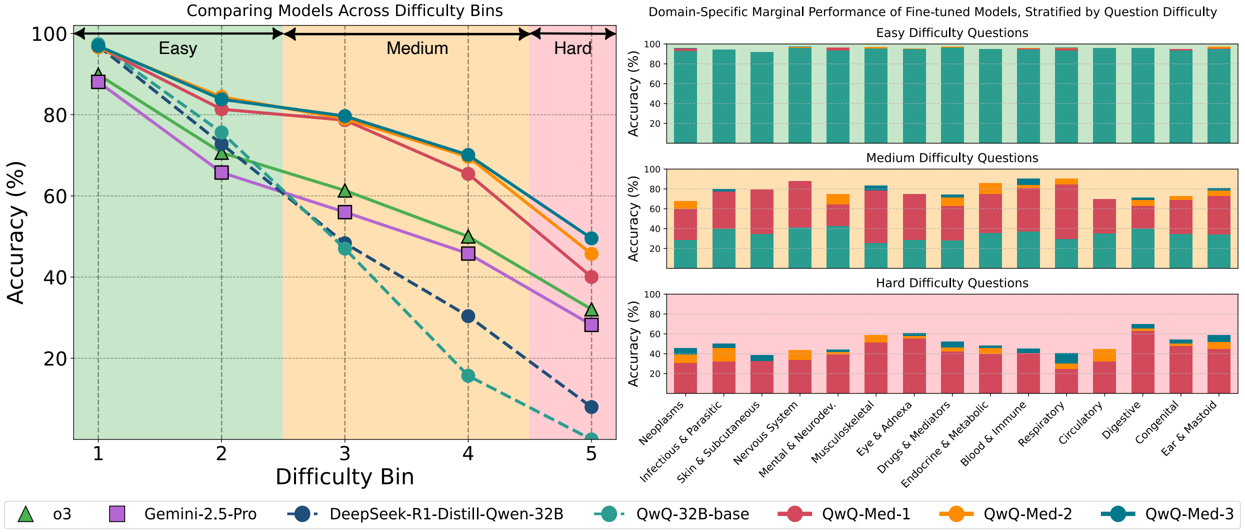


Figure 8: **Performance across task difficulty bins on ICD-Bench.** We construct a task-difficulty estimator using the pass@1 rate of the base model. Left: All models show declining accuracy with increasing difficulty, validating our pass@1-based difficulty estimator. Curriculum-tuned models exhibit improved robustness, with widening gains on increasingly challenging tasks. Right: Marginal gains across the curriculum tiers highlight that deep KG path exposure (QwQ-Med-3) is essential for solving the hardest questions, where the base model fails entirely.

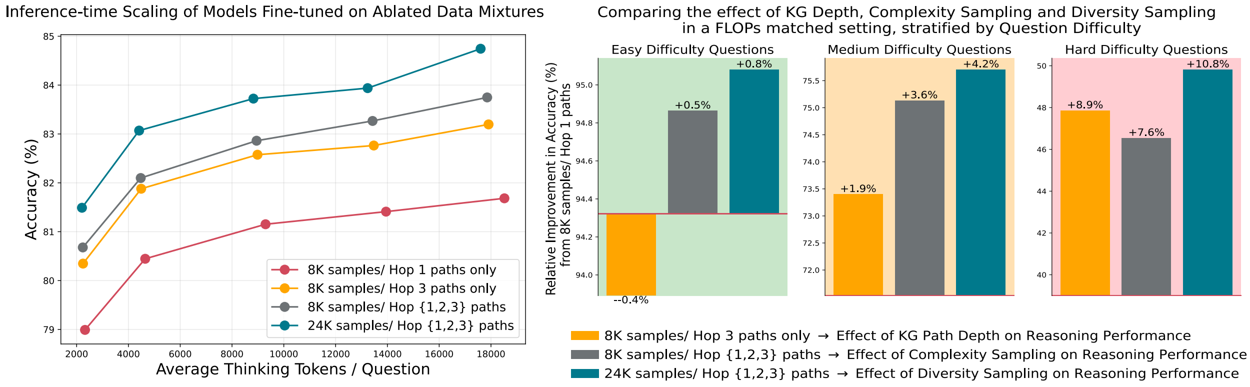


Figure 9: **Disentangling the effects of KG path depth, complexity sampling, and diversity sampling on curriculum-tuned model performance.** Left: Performance improves with increased thinking tokens as curricula incorporate deeper paths (3-hop), balanced path-length sampling, and greater diversity, demonstrating the additive benefits of each curation step. Right: Relative accuracy gains over a single-hop baseline stratified by task difficulty. Deeper paths are most helpful for challenging questions, balanced curricula are optimal for medium tasks, while easy tasks benefit most from diverse and balanced exposure.

and structured reasoning by diagnosing their generated thinking traces for alignment with the KG paths used to construct the question. This enables us to understand failure modes by disentangling errors due to inadequate recall from those arising due to erroneous inference.

(S5) Setup: For each ICD-Bench task, we verbalize the ground-truth KG path into individual hop-level premises. An LLM judge independently evaluates whether each hop is explicitly delineated in the model's reasoning trace, enabling partial credit for alignment. We report recall as the fraction of hops utilized and reasoning efficacy as the performance accuracy. In Fig. 10, we stratify ICD-Bench by KG hop lengths and report both metrics across hop-wise subsets. We observe:

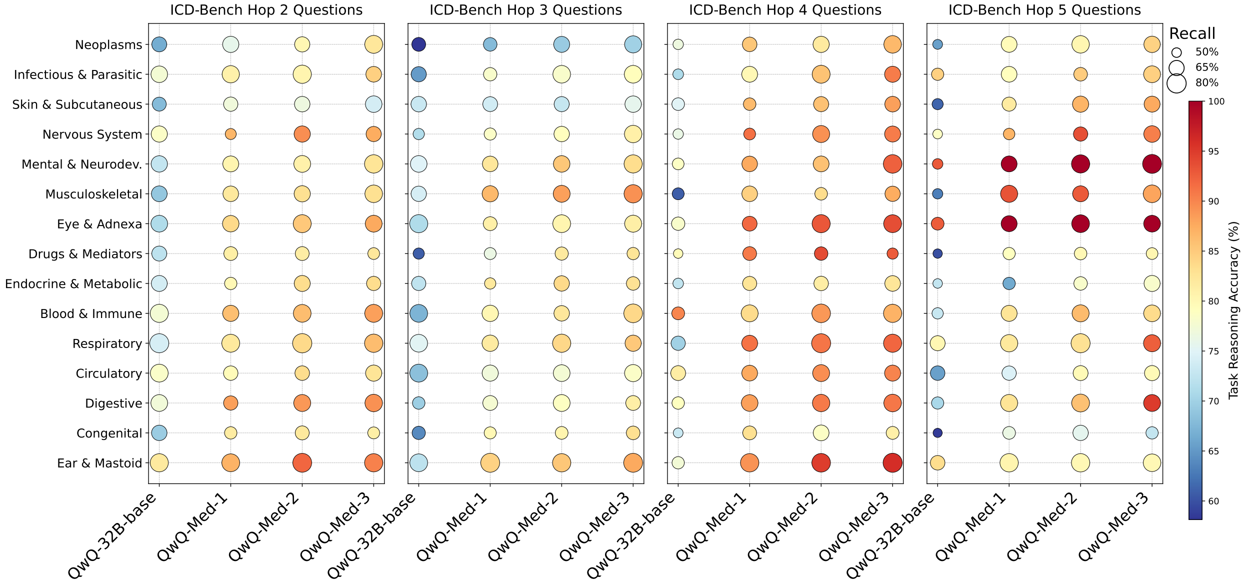


Figure 10: **Disentangling recall and reasoning performance across ICD-Bench tasks stratified by KG path length.** Dot size denotes recall (fraction of path entities recovered in the reasoning trace) and saturation reflects reasoning accuracy. Curriculum-tuned models demonstrate higher recall and effective reasoning across all hop levels, indicating successful use of KG primitives for reasoning. In contrast, the base model often retrieves relevant facts for shorter hop questions but fails to reason over them, revealing a surface-level understanding of the domain.

(O5.1) Curriculum-tuned models effectively utilize recalled paths. As shown in Fig. 10, curriculum-tuned models produce larger, more saturated dots across hop levels, reflecting both strong recall and effective reasoning. This indicates that these models are not merely retrieving KG primitives but effectively leveraging them to perform multi-step inference. Conversely, we also observe smaller, desaturated dots in certain categories (e.g., Drugs and Mediators), where even curriculum-tuned models struggle to reason correctly due to insufficient recall. These failure points underscore the importance of reliable recall for downstream reasoning, suggesting that model performance could further improve with more diverse training examples.

(O5.2) Base model can recall but fails to reason over retrieved knowledge. In contrast, the base model shows a notable disconnect between recall and reasoning. On two-hop and even some three-hop tasks, it retrieves relevant KG hops at moderate rates, but its reasoning accuracy remains significantly lower. This highlights a failure mode that indicates the base model possesses surface knowledge relevant to the task, yet struggles to integrate it into a coherent reasoning trace. On questions requiring longer reasoning paths (e.g., four- and five-hop chains), both recall and reasoning degrade sharply, indicating that the base model lacks the structural inductive biases to generalize beyond shallow retrieval.

6.6 Evaluation on Medical QA Benchmarks beyond the Original KG

To evaluate generalization beyond the scope of the KG, we assess our strongest model, QwQ-Med-3, on a suite of established medical QA benchmarks. Collectively, these datasets span a range of subdomains that provide a comprehensive evaluation for both robustness and generalization.

(S6) Setup: We benchmark the performance of QwQ-Med-3 against state-of-the-art open-source models: medical models like Meerkat (Kim et al., 2025) and MedGemma (Google, 2025), and general reasoning models like Deepseek-Distill-Qwen, Qwen 3 (Yang et al., 2025a), and Sky-T1 (Team, 2025a). We compare them across four widely-used benchmarks: MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), MMLU-Med subset (Hendrycks et al., 2020), and PubMedQA (Jin et al., 2019). Each model is evaluated under identical precision settings and standard accuracy metrics (See Table 3 in Appendix G). The results are summarized in Table 1.

(O6.1) Curriculum-tuned models reliably transfer acquired KG primitives. Our curriculum-tuned model, QwQ-Med-3, demonstrates competitive performance on external benchmarks, suggesting that the bottom-up primitives acquired through KG-grounded training can also be utilized for tasks beyond the original curriculum. However, these benchmarks, derived from medical board-style questions, primarily assess factual recall rather than structured reasoning. While such recall is essential, our curriculum tuning enables models to go beyond isolated facts, extending this knowledge into structured reasoning by learning to compose across facts. In contrast, baseline models, despite performing reasonably well on recall-based tasks, struggle to extend their recall capability effectively to more compositional reasoning, revealing the limits of their surface-level understanding.

Model	Size	# Training Examples	MedQA USMLE	PubMed QA	Med MCQA	MMLU Med-Subset
			Open-Source Models			
MedGemma[1]	27B	–	60.49	70.40	63.42	78.46
Meerkat	70B	441K	78.95	77.40	68.42	83.74
Deepseek-R1-Distill-Qwen	32B	800K	74.31	76.00	62.23	85.08
Qwen3	32B	–	64.34	65.40	55.63	69.53
Sky-T1	32B	17K	70.78	67.40	63.52	84.91
QwQ	32B	–	85.62	71.00	69.26	90.46
QwQ/parallel-scaling	32B	–	**87.09**	78.02	71.62	91.56
			Our Curriculum-Tuned Model			
QwQ-Med-3	32B	24K	82.72	76.00	71.03	90.64
QwQ-Med-3/parallel-scal.	32B	24K	85.39	**78.19**	**73.25**	**92.90**

Table 1: **Benchmarking our curriculum-tuned model against open-source models.** The bottom four rows compare the base QwQ model with our curriculum-tuned variant, including results obtained under inference-time scaling of the base and fine-tuned model. Best performance on a benchmark is highlighted in **bold**, with second-best performance underlined. Expanded results are presented in Table 4 in Appendix G.

7 Related Work

Reasoning with LLMs and KGs. Despite excelling on many natural language tasks, LLMs often struggle with complex reasoning and lack in-depth knowledge, often hallucinating facts in critical domains (Yu et al., 2025; Longwell et al., 2024). To mitigate this, researchers have integrated KGs as structured external sources to improve an LLM's reasoning and factual recall (Yasunaga et al., 2021; Su et al., 2024). In the medical domain, UMLS-based KGs have improved clinical and diagnostic reasoning, with adapter-based approaches effectively injecting UMLS knowledge into biomedical QA models (Gao et al., 2024a; Bodenreider, 2004; Park et al., 2023a; Schmidgall et al., 2024). Longitudinal studies have explored the use of smaller models, such as Graph Neural Networks or Long Short-Term Memory, to augment a larger model's reasoning (Chen et al., 2023; Zhang et al., 2022; Liu et al., 2024). Although these methods have achieved incremental success, their performance remains constrained due to the reliance on smaller, less powerful models. Contrary to the existing techniques that use Graph Retrieval-Augmented Generation to build an LLM+KG pipeline and rely on static external retrievers to obtain current facts (Liang et al., 2024), we propose grounding the model in

---

[1] Reported accuracy scores in the model card for MedGemma are different, using inference-time scaling, the details of which are undocumented.

intricate relationships during the pre-/post-training phase itself. Integration of knowledge bottom-up paves the way to building of superintelligent models capable of complex multi-hop reasoning.

Synthetic Data Curation for Instruction Tuning. Aligning LLMs with complex tasks via instruction tuning is limited by the scarcity of high-quality, human-annotated instruction-response pairs. To overcome this, researchers curate synthetic datasets with powerful base models (Zhang et al., 2023; Wang et al., 2022). Chen et al. (2024) illustrate this approach by introducing a GPT-4o-generated medical chain-of-thought (CoT) dataset with over 20K QA pairs; however, the dataset raises concerns about potential hallucinations in critical domains. Hybrid techniques, including a subset of expert-written examples, have achieved notable gains in domain-specific performance (Nazar et al., 2025; Zhang et al., 2023). To further ensure the quality of synthetic data samples, recent work leverages structured, verifiable sources, such as textbooks and excerpts, to generate CoT reasoning chains for existing questions (Kim et al., 2025) to boost the accuracy of small LMs. Tatarinov et al. (2025) use a query-based method (SPARQL) to extract QA pairs from a manually-annotated KG at multiple complexity levels for long-context tasks. In addition to a QA pair, our method leverages KGs to extract explicit reasoning paths towards the correct answer, thereby enhancing the quality of the generated dataset and imposing an inherent curriculum based on path length.

Curriculum Learning. Inspired by human pedagogy, Curriculum Learning (CL) is a training paradigm where models are progressively exposed to training examples arranged from easy to hard during pretraining to facilitate more effective learning (Bengio et al., 2009; Soviany et al., 2022). Early theoretical work demonstrated that difficulty-based ordering yields faster convergence and better performance on downstream tasks (Wang et al., 2024; Bengio et al., 2009). Just as medical students must learn anatomy and physiology before they can diagnose and treat complex diseases, a neural network must learn simple one-hop relational triples before complex multi-hop reasoning. Recent studies apply CL to LLMs, showing that exposure to incrementally harder questions enhances reasoning and instruction following (Ma et al., 2025; Luo et al., 2024; Thawakar et al., 2025). Notably, even small LMs can exhibit emergent multi-step reasoning when trained on carefully constructed curricula (Fu, 2025; Nair et al., 2024). We leverage the KG to generate training questions of increasing difficulty, using the number of hops as a proxy for complexity and conduct extensive experiments using diverse training recipes and scaling test-time compute to examine the role of CL in enabling reasoning depth and generalization in LMs (see Section 6).

Additional Related Work is presented in Appendix H.

8 Discussion

From Neural Abstractions to Data Abstractions. Our experimental findings underscore the importance of directly integrating reasoning traces derived from domain-specific primitives into the training data, rather than relying on the LM representations to learn them from examples that implicitly utilize, but do not make explicit, the underlying structure. We demonstrate the efficacy of this principle in the context of medicine, where a reliably curated KG readily provides bottom-up abstractions for synthesizing reasoning traces. Several recent works have also curated high-quality data using domain-specific abstractions from formal languages (Lin et al., 2025; Chervonyi et al., 2025), advanced examination questions (Muennighoff et al., 2025), and scientific forums (Yin et al., 2025). As we saturate the usage of Internet text for training data (Villalobos et al., 2024), designing domain-specific data abstractions that can seamlessly interface with natural language to synthesize high-quality training tasks is a promising direction.

Training/Inference Energy Cost Reduction. LLMs incur exorbitant energy costs during both training and inference. However, since the LMs that are fine-tuned for superintelligence can potentially be much smaller, their fine-tuning and inference energy costs can also be substantially reduced. Relying on a domain-specific architecture trained on abstracted data scaffolds to elicit high-quality reasoning, as opposed to a large architecture trained on unstructured Internet text, offers other efficiencies, such as requiring fewer inference tokens to achieve superintelligent expertise.

Bottom-up Primitives as Verifiable Rewards. Recent advances in reinforcement learning (RL) with LLMs have demonstrated success when guided by verifiable rewards (DeepSeekAI, 2025; Lambert et al., 2024; Yue et al., 2025), enabling significant strides in reasoning. While our current approach relies on SFT over full

KG paths to instill structured reasoning, the same setup can be naturally reframed as an RL problem. In this view, each KG primitive along a path functions as a localized verifier, providing a dense reward signal whenever the model correctly recalls or traverses a valid relational edge. This transforms the KG into a fully simulatable training environment, where reasoning agents can be optimized not only for end-task correctness but also for intermediate trace fidelity. Such a paradigm opens promising avenues for training superintelligent systems in domains where high-quality, bottom-up abstractions enable precise reward shaping.

Scalable Neurosymbolic Reasoning. A persistent criticism of deep learning systems is their inability to perform robust symbolic reasoning, as they primarily perform high-dimensional statistical interpolation over training distributions rather than explicitly represent and manipulate structured knowledge, causal relationships, or compositional rules of a domain (Marcus, 2025). Exact symbolic inference over large graphs is computationally intractable, and prior neurosymbolic systems, which rely on graph search or logic engines, have struggled to scale beyond toy domains. Our approach offers a scalable alternative by *neuralizing* symbolic reasoning primitives. KG paths encode discrete structure, while curriculum tuning on tasks derived from these paths teaches an LM to internalize compositional reasoning over the graph without explicitly enumerating all paths. This results in a differentiable approximation to symbolic reasoning that remains grounded in verifiable primitives, enabling tractable multi-hop inference within the reasoning trace itself.

Artificial General Intelligence (AGI) as Recursively Composable Bottom-Up Superintelligences. The dominant approach to AGI (Kokotajlo et al., 2025) today centers around scaling large monolithic architectures on domain-agnostic corpora to serve as a universal reasoning substrate across a broad spectrum of tasks. Our work lays the foundation for an alternative perspective in which general intelligence is an emergent property of a modular system of interacting superintelligent agents (Minsky, 1986; Andreas et al., 2017; Khot et al., 2023). In this imagined system, each agent can (1) specialize in a domain by learning from domain-specific abstractions (e.g., KGs) and (2) learn to communicate or hand off subproblems to adjacent specialists, forming a collaborative mesh of expertise. At inference time, complex tasks can then be decomposed into subtasks aligned with these specialized agents, with their outputs recursively composed along the agent-level compute graph to produce a coherent solution. This compositional model of AGI will require engineering domain-specific verifiable primitives that are functionally *local* to the agent as well as simulatable environments that allow *global* interactions to emerge from *local* primitives.

Limitations. While our work demonstrates promising results based on using a KG as a scaffold for deriving structured reasoning data, several constraints remain. First, the KG can be utilized beyond training to learn process reward models (PRM) (Lightman et al., 2023) from KG primitives and significantly improve inference-time scaling with PRM-guided search (Snell et al., 2024). Second, the underlying KG, despite providing a reliable structure over domain primitives, contains a closed vocabulary that constrains the conceptual coverage of the learned data abstractions. This limitation could be addressed by rigorously curating dense and high-quality KGs that cover diverse concepts. Third, we limit our focus to generating closed-ended multiple choice questions tasks. A significant challenge lies in being able to generate open-ended tasks from a KG that can be reliably transferred to real-world use cases (Microsoft AI, 2025). Fourth, our difficulty heuristic utilizes oracle answers to estimate task difficulty. Reliably learning a model-based difficulty metric without ground truth answers can be useful. Finally, we demonstrate the efficacy of our method in medicine where a reliable KG is available and its generalizability to other domains (e.g., law, banking), that lack canonical KGs or standardized abstractions, remains to be fully validated.

9 Conclusion

We introduced a novel task-synthesis framework that traverses structured paths on a KG to generate reasoning tasks that directly abstract domain-specific primitives. We also introduced ICD-Bench, a new evaluation suite designed to quantify domain-specific reasoning abilities over diverse medical domains. Using our approach, we curated a bottom-up curriculum of $24,000$ medical reasoning tasks and fine-tuned QwQ-32B on our dataset, resulting in our curriculum-tuned QwQ-Med-3 model. This model outperforms other reasoning baselines across ICD-Bench and performs comparably to them on other established benchmarks.

References

Samuel Joseph Amouyal, Tomer Wolfson, Ohad Rubin, Ori Yoran, Jonathan Herzig, and Jonathan Berant. QAMPARI: An Open-domain Question Answering Benchmark for Questions with Many Answers from Multiple Paragraphs. *CoRR*, abs/2205.12655, 2023.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural Module Networks. *CoRR*, abs/1511.02799, 2017.

Anthropic. Introducing Claude 4. Anthropic News, 2025. URL `https://www.anthropic.com/news/claude-4`. Accessed via Anthropic website.

Anthropic. Claude Code: Deep Coding at Terminal Velocity. Anthropic Engineering Blog, 2025. URL `https://www.anthropic.com/claude-code`. Agentic Coding Assistant Integrating with GitHub/GitLab and IDEs.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4825 LNCS, 2007.

Yoshua Bengio, Jerome Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. *ACM International Conference Proceeding Series*, 382, 2009. doi: 10.1145/1553374.1553380.

Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-Thought: Scaling Test-Time Compute for Enhancing LLM Reasoning. *CoRR*, abs/2412.09078, 2024.

Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32:D267–D270, 2004. doi: 10.1093/nar/gkh061.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, and Myra Deng et al. Genome Modeling and Design Across All Domains of Life with Evo 2. *bioRxiv*, abs/2025.02.18.638918, 2025. doi: 10.1101/2025.02.18.638918.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. Language Models are Few-Shot Learners. *CoRR*, abs/2005.14165, 2020.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision. *CoRR*, abs/2312.09390, 2023.

Susan Carey. *The Origin of Concepts.* Oxford University Press, New York, 2009.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. *CoRR*, abs/2412.18925, 2024.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Toward Subgraph-Guided Knowledge Graph Question Generation with Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):12706–12717, Apr. 2023.

Yuri Chervonyi, Trieu H. Trinh, Miroslav Olák, Xiaomeng Yang, Hoang Nguyen, Marcelo Menegali, Junehyuk Jung, Vikas Verma, Quoc V. Le, and Thang Luong. Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2. *CoRR*, abs/2502.03544, 2025.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems. *CoRR*, abs/2505.11831, 2025.

DeepSeek-AI. DeepSeek-V3 Technical Report. *CoRR*, abs/2412.19437, 2024.

DeepSeekAI. DeepSeekR1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *CoRR*, abs/2501.12948, 2025.

Xiang Deng, Ahmed Hassan Awadallah, Christopher Meek, Oleksandr Polozov, Huan Sun, and Matthew Richardson. Structure-Grounded Pretraining for Text-to-SQL. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia. In *The Semantic Web ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 2630, 2019, Proceedings, Part II*, pp. 6978, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-30795-0. doi: 10.1007/978-3-030-30796-7_5. URL `https://doi.org/10.1007/978-3-030-30796-7_5`.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B. Tenenbaum. DreamCoder: Growing Generalizable, Interpretable Knowledge with Wake-sleep Bayesian Program Learning. *CoRR*, abs/2006.08381, 2020.

Giorgio Franceschelli and Mirco Musolesi. Creative Beam Search: LLM-as-a-Judge For Improving Response Generation. *CoRR*, abs/2405.00099, 2024.

Xiang Fu. Can an Easy-to-Hard Curriculum Make Reasoning Emerge in Small Language Models? Evidence from a Four-Stage Curriculum on GPT-2. *CoRR*, abs/2505.11643, 2025.

Yanjun Gao, Ruizhe Li, Emma Croxford, Samuel Tesch, Daniel To, John Caskey, Brian W. Patterson, Matthew M. Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction. *medRxiv*, 2024a. doi: 10.1101/2023.11.24.23298641.

Zitian Gao, Boye Niu, Xuzheng He, Haotian Xu, Hongzhang Liu, Aiwei Liu, Xuming Hu, and Lijie Wen. Interpretable Contrastive Monte Carlo Tree Search Reasoning. *CoRR*, abs/2410.01707, 2024b.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What Can Transformers Learn In-Context? A Case Study of Simple Function Classes. *CoRR*, abs/2208.01066, 2023.

Google. Sharing Google's Med-PaLM 2 Medical Large Language Model, or LLM | Google Cloud Blog, 2024. URL `https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model`.

Google. MedGemma Model Card | Health AI Developer Foundations | Google for Developers, 2025. URL `https://developers.google.com/health-ai-developer-foundations/medgemma/model-card`.

Google Deepmind. AI Achieves Silver-Medal Standard Solving International Mathematical Olympiad Problems - Google DeepMind, 2024. URL `https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/`.

Google DeepMind. AlphaGenome: AI for Better Understanding the Genome. Google DeepMind Blog, Jun. 2025a. URL `https://deepmind.google/discover/blog/alphagenome-ai-for-better-understanding-the-genome/`.

Google DeepMind. Gemini 2.5 Flash Model Card, 2025b. URL `https://deepmind.google/models/gemini/flash/`.

Google DeepMind. Gemini 2.5 Pro Model Card, 2025c. URL `https://deepmind.google/models/gemini/pro/`.

Google Gemini Team. Gemini Deep Research: Your Personal Research Assistant. Google Gemini Website, 2024. URL `https://gemini.google/overview/deep-research/`. Agentic Research Feature Using Large Context Window and Search.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, and Katherine and Chou et al. Towards an AI Co-Scientist. *CoRR*, abs/2502.18864, 2025.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerating the Science of Language Models. *CoRR*, abs/2402.00838, 2024.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. In *Proceedings of the Web Conference 2021*, WWW 21, pp. 34773488. ACM, April 2021. doi: 10.1145/3442381.3449992. URL `http://dx.doi.org/10.1145/3442381.3449992`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *In Proceedings of the 9th International Conference on Learning Representations*, 2020.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E. Baranzini. Systematic Integration of Biomedical Knowledge Prioritizes Drugs for Repurposing. *eLife*, 6, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models. *CoRR*, abs/2203.15556, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *CoRR*, abs/2106.09685, 2021.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33, 2022. doi: 10.1109/TNNLS.2021.3070843.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of Experts. *CoRR*, abs/2401.04088, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei Hung Weng, Hanyi Fang, and Peter Szolovits. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences (Switzerland)*, 11, 2020.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2567–2577, 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature*, 596:583–589, 2021.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed Prompting: A Modular Approach for Solving Complex Tasks. *CoRR*, abs/2210.02406, 2023.

Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C. Rindflesch. SemMedDB: A PubMed-Scale Repository of Biomedical Semantic Predications. *Bioinformatics*, 28, 2012.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. Small Language Models Learn Enhanced Reasoning Skills from Medical Textbooks. *NPJ Digital Medicine*, 8, 2025.

Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. AI 2027: We Predict That the Impact of Superhuman AI Over the Next Decade Will Be Enormous, Exceeding That of the Industrial Revolution, Apr. 2025. URL `https://ai-2027.com/ai-2027.pdf`.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building Machines That Learn and Think Like People. *CoRR*, abs/1604.00289, 2016.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester J. V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Øyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing Frontiers in Open Language Model PostTraining. *CoRR*, abs/2411.15124, 2024.

Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, Xinwang Liu, Fuchun Sun, and Kunlun He. A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multimodal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9456–9478, Jan. 2024. doi: 10.1109/TPAMI.2024.3417451.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. *CoRR*, abs/2305.20050, 2023.

Yong Lin, Shange Tang, Bohan Lyu, Jiayun Wu, Hongzhou Lin, Kaiyu Yang, Jia Li, Mengzhou Xia, Danqi Chen, Sanjeev Arora, and Chi Jin. Goedel-Prover: A Frontier Model for Open-Source Automated Theorem Proving. *CoRR*, abs/2502.07640, 2025.

Guangyi Liu, Yongqi Zhang, Yong Li, and Quanming Yao. Explore then Determine: A GNN-LLM Synergy Framework for Reasoning over Knowledge Graph. *CoRR*, abs/2406.01145, 2024.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *CoRR*, abs/2307.03172, 2023.

Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1B LLM Surpass 405B LLM? Rethinking Compute-Optimal Test-Time Scaling. *CoRR*, abs/2502.06703, 2025.

Jack B. Longwell, Ian Hirsch, Fernando Binder, Galileo Arturo Gonzalez Conchas, Daniel Mau, Raymond Jang, Rahul G. Krishnan, and Robert C. Grant. Performance of Large Language Models on Medical Oncology Examination Questions. *JAMA Network Open*, 7, 2024.

Kangyang Luo, Zichen Ding, Zhenmin Weng, Lingfeng Qiao, Meng Zhao, Xiang Li, Di Yin, and Jinlong Shu. Let's Be Self-Generated via Step by Step: A Curriculum Learning Approach to Automated Reasoning with Large Language Models. *CoRR*, abs/2410.21728, 2024.

Xuetao Ma, Wenbin Jiang, and Hua Huang. Problem-Solving Logic Guided Curriculum In-Context Learning for LLMs Complex Reasoning. *CoRR*, abs/2502.15401, 2025.

Chaitanya Malaviya, Peter Shaw, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. QUEST: A Retrieval Dataset of Entity-Seeking Queries with Implicit Set Operations. *CoRR*, abs/2305.11694, 2023.

Daniel J. Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, and Marco Selvi. Faster Sorting Algorithms Discovered Using Deep Reinforcement Learning. *Nature*, 2023.

Gary Marcus. Generative AI's Crippling and Widespread Inability to Reason. `https://garymarcus.substack.com/p/generative-ais-crippling-and-widespread`, 2025. Accessed: 2025-08-30.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, Ekin Dogus Cubuk, et al. Scaling Deep Learning for Materials Discovery. *Nature*, 614:1234–1240, 2023. doi: 10.1038/s41586-023-06735-9.

Meta AI. The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation. Meta AI Blog, Apr. 2025. URL `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

Microsoft AI. The Path to Medical Superintelligence, 2025. URL `https://microsoft.ai/new/the-path-to-medical-superintelligence/`.

Marvin Minsky. *The Society of Mind*. Simon & Schuster, New York, 1986. ISBN 0-671-60740-5. First Comprehensive Presentation of the "Society of Mind" Theory.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple Test-Time Scaling. *CoRR*, abs/2501.19393, 2025.

Marwa Nair, Kamel Yamani, Lynda Said Lhadj, and Riyadh Baghdadi. Curriculum Learning for Small Code Language Models. *CoRR*, abs/2407.10194, 2024.

Wojciech Nazar, Grzegorz Nazar, Aleksandra Kamińska, and Ludmila Danilowicz-Szymanowicz. How to Design, Create, and Evaluate an Instruction-Tuning Dataset for Large Language Model Training in Health Care: Tutorial From a Clinical Perspective. *Journal of Medical Internet Research*, 27, 2025. doi: 10.2196/70481.

Harsha Nori, Nicholas King, Scott Mayer Mckinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on Medical Challenge Problems. *CoRR*, abs/2303.13375, 2023.

Alexander Novikov, Ngan Vu, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Z. Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. AlphaEvolve: A Coding Agent for Scientific and Algorithmic Discovery. *CoRR*, abs/2506.13131, 2025.

OpenAI. Hello GPT-4o, 2024. URL `https://openai.com/index/hello-gpt-4o/`. System Card and Technical Overview.

OpenAI. Introducing Codex. OpenAI Blog, 2025a. URL `https://openai.com/index/introducing-codex/`. Research preview; Codex-1 model.

OpenAI. Introducing Deep Research. OpenAI Blog, February 2025b. URL `https://openai.com/index/introducing-deep-research/`. Launch of the Deep Research Feature Within ChatGPT, Powered by a Specialized Version of the o3 Model.

OpenAI. Introducing OpenAI o3 and o4-mini, 2025c. URL `https://openai.com/index/introducing-o3-and-o4-mini/`.

World Health Organization. *International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)*. World Health Organization, 1992. URL `https://icd.who.int/browse10/2019/en`.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA : A Large-Scale Multi-Subject Multi-Choice Dataset for Medical Domain Question Answering. *In Proceedings of Machine Learning Research*, 174:248–260, Apr. 2022.

Hyeryun Park, Jiye Son, Jeongwon Min, and Jinwook Choi. Selective UMLS Knowledge Infusion for Biomedical Question Answering. *Scientific Reports*, 13, 2023a.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith R. Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. *CoRR*, abs/2304.03442, 2023b.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. Safety Alignment Should Be Made More Than Just a Few Tokens Deep. *CoRR*, abs/2406.05946, 2024.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? *CoRR*, abs/2304.15004, 2023.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Reis, Jeffrey Jopling, and Michael Moor. AgentClinic: A Multimodal Agent Benchmark to Evaluate AI in Simulated Clinical Environments. *CoRR*, abs/2405.07960:2025–2030, 2024.

Amit Singhal. Introducing the Knowledge Graph: Things, Not Strings, 2012. URL `https://blog.google/products/search/introducing-knowledge-graph-things-not/`.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Model Parameters. *CoRR*, abs/2408.03314, 2024.

Petru Soviany, Tudor Radu, Paolo Rota, and Nicu Sebe. Curriculum Learning: A Survey. *International Journal of Computer Vision*, 130, 2022. doi: 10.1007/s11263-022-01611-x.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA. *CoRR*, abs/2410.04660, 2024.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive Architectures for Language Agents. *CoRR*, abs/2309.02427, 2023.

Nikita Tatarinov, Vidhyakshaya Kannan, Haricharana Srinivasa, Arnav Raj, Singh Anand, Varun Singh, Aditya Luthra, Ravij Lade, Agam Shah, and Sudheer Chava. KG-QAGen: A Knowledge-Graph-Based Framework for Systematic Question Generation and Long-Context LLM Evaluation. *CoRR*, abs/2505.12495v1, 2025.

Gemini Team, DeepMind, and Google Research. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. *CoRR*, abs/2403.05530, 2024.

NovaSky Team. Sky-T1: Fully Open-Source Reasoning Model with o1-Preview Performance in $450 Budget, 2025a. URL `https://novasky-ai.github.io/posts/sky-t1`.

Qwen Team. QwQ-32B: Embracing the Power of Reinforcement Learning, 2025b. URL `https://qwenlm.github.io/blog/qwq-32b/`.

Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331, 2011. doi: 10.1126/science.1192788.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. LlamaV-o1: Rethinking Step-by-step Visual Reasoning in LLMs. *CoRR*, abs/2501.06186, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *CoRR*, abs/2305.04388, 2023.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data. *CoRR*, abs/2211.04325, 2024.

Xin Wang, Yuwei Zhou, Hong Chen, and Wenwu Zhu. Curriculum Learning: Theories, Approaches, Applications, Tools, and Future Directions in the Era of Large Language Models. *Companion Proceedings of the ACM Web Conference*, 2024. doi: 10.1145/3589335.3641257.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *CoRR*, abs/2203.11171, 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 2022.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *CoRR*, abs/2206.07682, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *CoRR*, abs/2201.11903, 2023.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *CoRR*, abs/1710.06481, 2018.

David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, Jan. 2018. doi: 10.1093/nar/gkx1037.

Zhaolong Wu, Abul Hasan, Jinge Wu, Yunsoo Kim, Jason P. Y. Cheung, Teng Zhang, and Honghan Wu. Chain-of-Thought (CoT) Prompting Strategies for Medical Error Detection and Correction. *Association for Computational Linguistics*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 Technical Report. *CoRR*, abs/2505.09388, 2025a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115, 2025b.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering. *In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Ming Yin, Yuanhao Qu, Ling Yang, Le Cong, and Mengdi Wang. Toward Scientific Reasoning in LLMs: Training from Expert Discussions via Reinforcement Learning. *CoRR*, abs/2505.19501, 2025.

Tao Yu, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong. GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing. *CoRR*, abs/2009.13845, 2021.

Tong Yu, Yongcheng Jing, Xikun Zhang, Wentao Jiang, Wenjie Wu, Yingjie Wang, Wenbin Hu, Bo Du, and Dacheng Tao. Benchmarking Reasoning Robustness in Large Language Models. *CoRR*, abs/2503.04550, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *CoRR*, abs/2504.13837, 2025.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. Med-PRM: Medical Reasoning Models with Stepwise, Guideline-verified Process Rewards. *CoRR*, abs/2506.11474, 2025.

Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. Subgraph Retrieval Enhanced Model for Multi-hop Knowledge Base Question Answering. *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1, 2022.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. AlpaCare: Instruction-tuned Large Language Models for Medical Application. *CoRR*, abs/2310.14558, 2023.

Yuyu Zhang, Xinshi Chen, Yuan Yang, Arun Ramamurthy, Bo Li, Yuan Qi, and Le Song. Efficient Probabilistic Logic Reasoning with Graph Neural Networks. *CoRR*, abs/2001.11850, 2020.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *CoRR*, abs/1709.00103, 2017.

Appendices

A Generating Grounded Question-Answering Tasks Using a KG

A.1 UMLS KG Details

We use a KG constructed by past work (Yasunaga et al., 2021) that integrates the Disease Database portion of the Unified Medical Language System (UMLS) (Bodenreider, 2004) and DrugBank (Wishart et al., 2018). The resultant KG contains 9,958 nodes and 44,561 edges. We utilized all edge relation types of the KG, with the exception of "belongs to the category of," "is a category," and "is a subtype of," to avoid generating tasks that test UMLS taxonomy-based recall. Fig. 11 shows the distributional statistics of the KG.

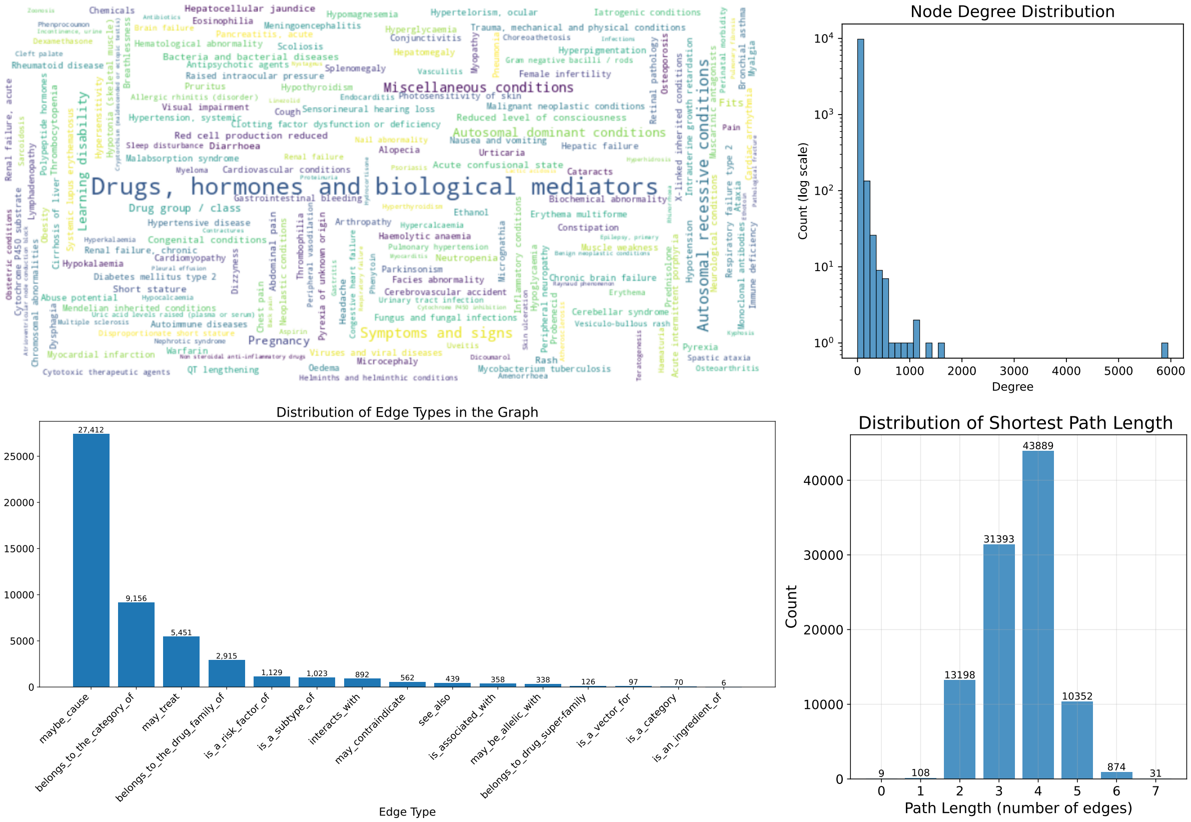


Figure 11: **Distributional statistics of the KG.** Top row: On the left, we visualize the top-200 nodes with the highest degree in a word cloud. On the right, we plot the histogram of node degrees. The KG is fairly sparse and the weight is significantly concentrated on single-degree nodes. Bottom: On the left, we show the breakdown of the edges across their relation types. On the right, we randomly sample 100,000 node pairs and measure the shortest path distance between them. Most nodes are ∼ 4 hops away from each other, as a result of the sparsity of the KG.

A.2 QA Generation Prompt

Prompt 1 details the template used to transform KG paths into QA pairs. Given a KG path, the template generates a reasoning task that links the initial entity to the final entity by traversing the intermediate relations.

Prompt 1: QA Generation from KG path Prompt Template

Create a medical examination question for advanced medical students that tests the relationship between {#insert source entity} and {# insert target entity}. The relationship is: {# insert KG path here}. The question should:

1. Be in multiple choice format (4 options)
2. Require clinical reasoning along the relationship
3. Include a brief clinical vignette
4. Not directly mention the relationship in the question stem
5. Have one clearly correct answer

Format:

```
<Question>
    [Clinical Vignette]
</Question>
<Options>
    A. [Option]
    B. [Option]
    C. [Option]
    D. [Option]
</Options>
<Answer>:
    [Correct Option Letter]
</Answer>
```

A.3 Task Generation Pipeline Details

Prompt 2 presents the template used to generate a thinking trace for a QA pair, conditioned on its corresponding KG path. Prompt 3 defines the template employed by the LLM grader to assess correctness. The grader evaluates the alignment between the answer, thinking trace, and KG context, using the full QA pair as input. The overall curriculum curation pseudo-code is presented in Algorithm 1.

Prompt 2: Thinking Trace Generation Prompt Template

Generate a detailed explanation for the question: {#insert question and options}
Use the following context {#insert KG path here}. The explanation should be:

1. Detailed and include all the steps leading to the answer.
2. You are to use the provided context to explain the relationship between the concepts.
3. Strictly do not mention that you are using a given context to generate the explanation.

Prompt 3: Correctness Filtering Prompt Template

You are a medical examiner. You are given a medical question along with an explanation and the answer. You have also been given a source context.

1. Judge whether the question and answer are logically correct and medically accurate, and follow the source. If there is an explanation, also evaluate whether the explanation follows from the source to reach the correct answer.
2. Respond with only "Yes" or "No".

Format your response exactly like this: “Correct: [Yes/No]”
Question: {# insert question and options here}

Explanation: {# insert thinking trace here}
Answer: {# insert answer option letter here}
Source Context: {# insert KG path here}

Algorithm 1: Curriculum Curation Pseudo-code

```
Input: Knowledge Graph $\mathcal{G}$, Max Path Length $N$, QA template $\mathcal{T}_{\text{QA}}$, Thinking Model Prompt $\mathcal{T}_{\text{think}}$,
       Correctness Filtering Prompt $\mathcal{T}_{\text{verify}}$, TotalSamples
Output: High-quality QA pairs with grounded thinking traces
Initialize node frequency table $\{f_i = 0\}$ for each $i \in \mathcal{G}$;
Initialize dataset $\mathcal{D} \leftarrow \emptyset$;
while size($\mathcal{D}$) != TotalSamples do
    Sample source node $h_0 \sim \text{InverseFreqSampling}(f_i)$ // Diversity Sampling;
    Sample path length $L \sim \text{Uniform}(\{1, \dots, N\})$ // Complexity Sampling;
    Sample path $p^L = (h_0, r_1, h_1, \dots, r_L, h_L)$ from $\mathcal{G}$ // KG Path Sampling;
    Generate $(q, A, \text{Options}) =$ Gemini-2.0-Flash($\mathcal{T}_{\text{QA}}(h_0, h_L, p^L)$) // QA Generation;
    // Quality Filtering;
    if Invalid formatting, distractors, or missing fields then
        continue;
    Generate trace $T =$ Gemini-2.5-Pro($\mathcal{T}_{\text{think}}(q, \text{Options}, p^L)$) // Thinking Trace Generation;
    // Correctness Filtering with Dual LLMs;
    $\text{verdict}_1 =$ Gemini-2.0-Flash($\mathcal{T}_{\text{verify}}(q, A, T, p^L)$);
    $\text{verdict}_2 =$ Qwen-2.5-72B($\mathcal{T}_{\text{verify}}(q, A, T, p^L)$);
    if $verdict_1$ == pass and $verdict_2$ == pass then
        Add $(q, A, \text{Options}, T)$ to $\mathcal{D}$;
        Increment frequency count $f_i$ for all nodes in $p^L$;
return $\mathcal{D}$
```

B UMLS-Curated Curriculum Details

Examples 3 and 4 illustrate generated samples of the thinking trace in our UMLS-curated curriculum, used to fine-tune our curriculum-tuned models. The traces are generated by `Gemini 2.5 pro` using Prompt 2. We summarize our dataset in Table 2 by aligning questions to the ICD taxonomy.

Example 3: Thinking trace from the UMLS-curated training curriculum

KG-path:

Au-Kline syndrome $\xrightarrow{\text{maybe cause}}$ Blepharoptosis $\xleftarrow{\text{maybe cause}}$ Turner syndrome $\xrightarrow{\text{maybe cause}}$ Cubitus valgus

Question:

A 16-year-old female presents to your clinic with a long-standing history of bilateral ptosis. On examination, you note mild developmental delay, subtle dysmorphic facial features, and a history of recurrent upper respiratory infections. She denies any muscle weakness or fatigability. Her height is in the 25th percentile for her age. You order genetic testing, and the results are pending. Based on the most likely underlying diagnosis, which of the following physical examination findings would you be MOST likely to identify?

Options:

A. Proximal muscle weakness and elevated creatine kinase B. Webbed neck and short stature C. Down-sloping palpebral fissures and epicanthal folds **D. Increased carrying angle of the elbows**

Gemini 2.5 Pro Thinking Trace:

Okay, let's break this down like we're figuring it out together for rounds.

1. **Analyze the Patient's Presentation:** We have a 16-year-old female with a few key things going on:
   - **Bilateral ptosis (droopy eyelids):** This has been there a long time.
   - **Mild developmental delay & subtle dysmorphic features:** Suggests a possible underlying genetic syndrome.
   - **Recurrent URIs:** Could be related to immune function or anatomical issues (like Eustachian tube problems).
   - **No muscle weakness/fatigability:** This makes things like myasthenia gravis or progressive muscular dystrophies less likely as the *primary* issue explaining the whole picture.
   - **Height at 25th percentile:** She's on the shorter side for her age, but not profoundly short. Still, short stature in the context of other features is significant.
2. **Consider the Differential Diagnosis based on the key features:** The combination of ptosis, short stature (relative), developmental delay, and dysmorphic features screams "genetic syndrome." Let's look at the options through this lens:
   - **A. Proximal muscle weakness and elevated CK:** This points towards a primary muscle disease (myopathy/dystrophy). While some syndromes with ptosis *can* have muscle involvement (like mitochondrial myopathies or myotonic dystrophy), the patient specifically *denies* weakness or fatigability, making this unlikely to be the *most* likely additional finding.
   - **C. Down-sloping palpebral fissures and epicanthal folds:** These are very characteristic features of Down syndrome (Trisomy 21). Down syndrome does involve developmental delay, dysmorphic features, and increased risk of infections. However, ptosis isn't typically listed as a *primary* or defining feature, although it can occur. Also, the question presents *ptosis* as a major complaint alongside the other features, suggesting it's central to the suspected diagnosis.

- **B. Webbed neck and short stature & D. Increased carrying angle of the elbows (Cubitus valgus):** Both of these are strongly associated with **Turner Syndrome (TS)**. Let's see if TS fits the overall picture:
  - **Cause:** Usually 45,X karyotype or mosaicism. Affects females.
  - **Ptosis:** Yes, ptosis can occur in Turner Syndrome.
  - **Short Stature:** This is a hallmark feature. While often below the 5th percentile, presentation can vary, and 25th percentile is certainly relatively short and fits.
  - **Developmental Delay:** Usually, intelligence is normal in TS, but specific non-verbal learning disabilities are common and might be described as "mild developmental delay" in some contexts.
  - **Dysmorphic Features:** TS has characteristic features like low posterior hairline, high-arched palate, micrognathia, and sometimes epicanthal folds (though maybe not typically down-sloping fissures). A webbed neck (pterygium colli) is classic (Option B). Skeletal anomalies like a broad chest and cubitus valgus (increased carrying angle at the elbow, Option D) are also very common.
  - **Recurrent Infections:** Recurrent otitis media is common in TS due to Eustachian tube dysfunction, which might be reported by the family as recurrent URIs.

3. **Choose the MOST Likely Finding:** So, Turner Syndrome seems like a very strong candidate diagnosis given the patient's features (female, ptosis, relative short stature, possible dysmorphism/DD). Both webbed neck (B) and increased carrying angle (D) are features of TS.
   - A webbed neck is a classic sign but isn't present in all individuals with TS (maybe 25–50%) and can be subtle.
   - An increased carrying angle (cubitus valgus) is a *very* common skeletal finding in TS, reported in over 50% (some sources say >70%) of individuals. It's often readily apparent on physical exam when checking the arms.
   - Therefore, between B and D, the increased carrying angle (cubitus valgus) is statistically more frequent and thus the *most likely* additional finding you'd identify on examination if the underlying diagnosis is indeed Turner Syndrome.
4. **Final Conclusion:** Based on the patient's constellation of symptoms pointing strongly towards Turner Syndrome as the most likely underlying diagnosis, an increased carrying angle of the elbows (cubitus valgus) is the most probable associated physical finding among the choices provided.

Example 4: Thinking trace from the UMLS-curated training curriculum

KG-path: Aripiprazole$\xrightarrow{\text{may treat}}$ Schizophrenia

Question:

A 24-year-old male with a history of psychosis presents to the emergency department with increasing agitation, disorganized speech, and paranoid ideation. He reports auditory hallucinations that command him to harm himself. His family confirms a previous diagnosis of a chronic mental illness and that he stopped taking his prescribed medication approximately 3 months ago due to feeling "like a zombie." Physical examination is unremarkable. After initial stabilization with a short-acting benzodiazepine, which of the following long-term medication strategies is MOST appropriate to address both the positive and negative symptoms while minimizing the risk of extrapyramidal side effects and significant weight gain?

Options:

A. Haloperidol decanoate injection every 4 weeks. B. Clozapine with weekly white blood cell count monitoring. C. Risperidone long-acting injection every 2 weeks. **D. Aripiprazole oral daily.**

Gemini 2.5 Pro Thinking Trace:

Okay, let's break down this question step-by-step to figure out the best long-term medication strategy for this patient.

1. **Understand the Patient and Situation:** We have a young adult male (24 years old) experiencing an acute psychotic episode (agitation, disorganized speech, paranoia, command hallucinations). This is likely an exacerbation of a known chronic mental illness, probably schizophrenia spectrum disorder given the symptoms and age. A key factor is that he *stopped* his previous medication because of side effects, describing feeling "like a zombie," which often implies sedation, cognitive slowing, or extrapyramidal symptoms (EPS) like Parkinsonism or akinesia. He's been stabilized short-term with a benzodiazepine, but now needs a long-term plan.
2. **Identify the Goals of Treatment:** The question specifically asks for a medication that:
   - Treats both positive symptoms (hallucinations, paranoia, disorganized thought) AND negative symptoms (which are common in chronic psychosis, like apathy, flat affect, alogia - though not explicitly described here, they are a target for treatment).
   - Minimizes the risk of EPS (like Parkinsonism, dystonia, akathisia, tardive dyskinesia). This is crucial given his previous experience ("zombie" feeling).
   - Minimizes the risk of significant weight gain (a common and problematic side effect of many antipsychotics).
3. **Evaluate the Options Based on Goals:**
   - **A. Haloperidol decanoate injection:** This is a long-acting injectable (LAI) form of a first-generation antipsychotic (FGA).
     - *Pros:* LAI helps with adherence, which is good since he stopped meds before. Effective for positive symptoms.
     - *Cons:* Haloperidol is notorious for causing EPS. It has limited effectiveness against negative symptoms. The "zombie" feeling could very well have been EPS from a previous similar drug, making this a poor choice for minimizing that specific risk.
     - *Conclusion:* Doesn't meet the criteria for minimizing EPS or treating negative symptoms well.
   - **B. Clozapine:** This is a second-generation antipsychotic (SGA), often considered the most effective.
     - *Pros:* Highly effective for both positive and negative symptoms, especially in treatment-resistant cases. Very low risk of EPS.
     - *Cons:* Requires rigorous weekly (initially) blood monitoring due to the risk of agranulocytosis (dangerously low white blood cells). It also carries a high risk of significant weight gain, sedation, constipation, and other side effects. It's typically reserved for patients who haven't responded adequately to at least two other antipsychotics.
     - *Conclusion:* While it has low EPS risk and treats both symptom types, the high risk of weight gain and the significant monitoring burden/safety concerns make it inappropriate as a first-line or second-line choice here. This patient hasn't necessarily failed other treatments yet.
   - **C. Risperidone long-acting injection:** This is an LAI form of an SGA.
     - *Pros:* LAI helps with adherence. Effective for positive symptoms and some effect on negative symptoms.
     - *Cons:* Among SGAs, risperidone has a relatively higher risk of causing EPS (especially at higher doses) and hyperprolactinemia (which can cause sexual dysfunction, gynecomastia). It's also associated with moderate-to-significant weight gain.
     - *Conclusion:* While better than haloperidol, it still carries a considerable risk for EPS and weight gain, which we want to minimize.
   - **D. Aripiprazole oral daily:** This is an SGA with a unique mechanism (D2 partial agonism).

  - *Pros:* Effective for positive symptoms and potentially beneficial for negative and cognitive symptoms. It has a *lower* risk of EPS compared to FGAs and many other SGAs like risperidone (though akathisia can occur). Crucially, it has a *lower* risk of significant weight gain and metabolic side effects compared to agents like clozapine, olanzapine, and risperidone. Aripiprazole is known to be generally less sedating for many patients.
  - *Cons:* It's an oral medication, so adherence could still be an issue. However, starting with an oral agent that has a better side-effect profile is a reasonable first step, addressing the *reason* he stopped his previous medication. If adherence fails despite better tolerability, an LAI version of aripiprazole exists.
  - *Conclusion:* This option provides the best balance. It targets positive and negative symptoms, has a lower risk profile for the specific side effects we're concerned about (EPS and weight gain), and directly addresses the likely reason for his previous non-adherence (poor tolerability). Aripiprazole is often considered a good choice when trying to balance efficacy with minimizing metabolic side effects and EPS.

4. **Final Decision:** Comparing the options, Aripiprazole (D) best fits the requirements laid out in the question: efficacy for positive/negative symptoms, minimized EPS risk, and minimized weight gain risk, making it the most appropriate long-term strategy for this patient given his history and presentation.

Category	Total Questions	Avg Thinking Tokens	Total Tokens	Most Frequent Entities (Top 5)
Certain infectious and parasitic diseases	2588	1128.03	2919330	Gram negative bacilli / rods, Meningoencephalitis, Labyrinthitis, Mycobacterium tuberculosis, Viral haemorrhagic fever
Neoplasms	1862	1076.20	2003880	Bronchogenic carcinoma, Melanoma, Hemangioma, Myeloma, Colorectal cancer
Blood/immunity disorders	2425	1173.51	2845761	Thrombocytopenia, Neutropenia, Red cell production reduced, Thrombophilia, Haemolytic anaemia
Endocrine/nutritional/metabolic	3387	1198.75	4060163	Hyperglycaemia, Hypoglycaemia, Enzymes, Obesity, Malabsorption syndrome
Mental and neurodevelopmental disorders	1824	1069.55	1950856	Learning disability, Sleep disturbance, Labyrinthitis, Involuntary muscular movements, Deliberate self harm
Nervous system diseases	4083	1118.38	4566330	Learning disability, Headache, Acute confusional state, Chronic brain failure, Cerebellar syndrome
Eye and adnexa diseases	1974	1091.62	2154860	Retinal pathology, Cataracts, Corneal opacity, Conjunctivitis, Eye pain
Ear and mastoid diseases	1064	1035.77	1102058	Sensorineural hearing loss, Conductive hearing loss, Tinnitus, Suppurative otitis media, Hearing loss
Circulatory system diseases	3213	1130.10	3631027	Respiratory failure type 2, QT lengthening, Pulmonary hypertension, Spastic ataxia, Cardiomyopathy
Respiratory system diseases	1663	1092.93	1817539	Respiratory failure type 2, Bronchial asthma, Breathlessness, Pneumonia, Cough
Digestive system diseases	2932	1108.66	3250596	Diarrhoea, Gastrointestinal bleeding, Renal failure (chronic), Pyrexia of unknown origin, Dysphagia
Skin and subcutaneous tissue diseases	6646	1129.23	7504844	Hepatocellular jaundice, Headache, Pruritus, Neutropenia, Hepatomegaly
Musculoskeletal/connective tissue diseases	4218	1128.49	4759982	Fits, Arthropathy, Muscle weakness, Rheumatoid disease, Myalgia
Congenital/chromosomal abnormalities	4065	1145.85	4657897	Microcephaly, Micrognathia, Deafness onychodystrophy syndrome, Cleft palate, Syndactyly
Drugs, hormones, and mediators	2163	1138.08	2461671	Cytochrome P450 substrate, Cytotoxic therapeutic agents, Dexamethasone, Cyclophosphamide, Dicoumarol

Table 2: **Summary of our generated curriculum:** We categorize each generated question into one or more ICD-10 categories by checking whether its KG path includes an entity from that category. For each category, we compute the total token count and the average length in the thinking traces, using the QwQ tokenizer. In addition, we identify the most frequent entities in each category to serve as representative keywords.

C Supervised Fine-Tuning Setup

We fine-tuned the off-the-shelf QwQ-32B reasoning model (Team, 2025b), which demonstrates strong performance across several medical QA benchmarks and exhibits broad understanding of medical concepts. During SFT, we injected the thinking trace of the model between `<think>` and `</think>` delimiters, followed by the final answer. The next-token prediction objective was applied only to the reasoning trace and the answer. We employed LoRA (Hu et al., 2021) with rank $= 16$ and $\alpha = 16$, using a batch size of 16 with gradient accumulation every two steps across 8×H100 GPUs. The learning rate followed a cosine schedule with a peak of 1e−5 and linear warmup over the first 5% of training steps. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 1e−4. We ran SFT for 8 epochs on the full dataset. For models trained on subsets, we adjusted the number of epochs to match the total training FLOPs of the full-scale setting. Fig. 12 visualizes the training dynamics of our fine-tuned models.

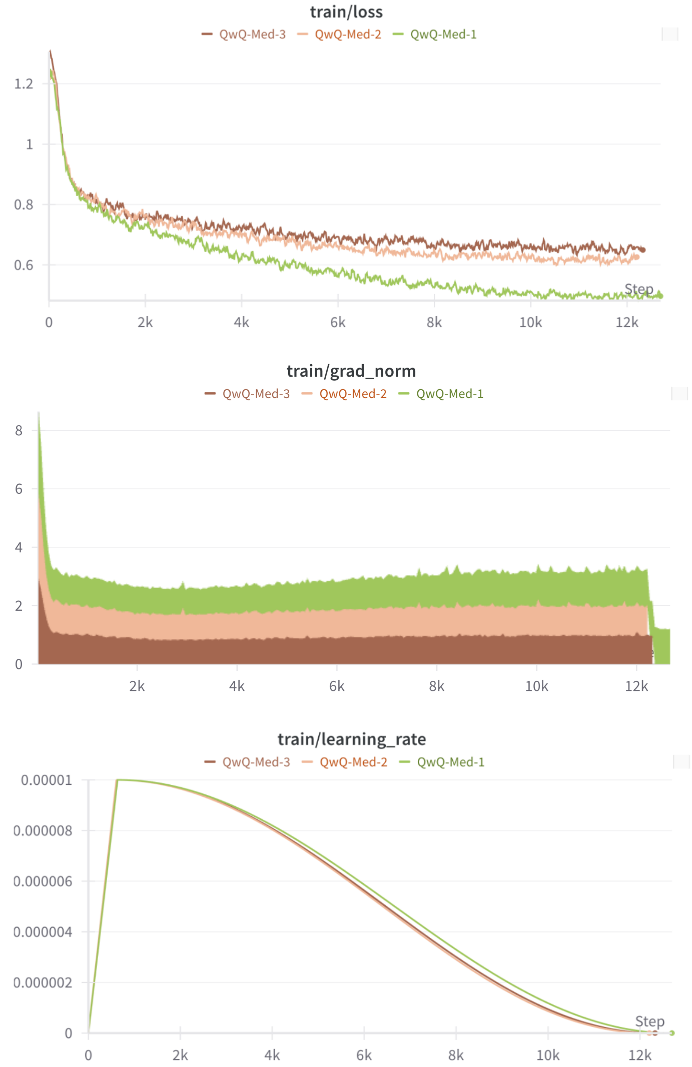


Figure 12: Training run details of our curriculum-tuned models.

D Additional Examples: Qualitative Comparison on ICD-Bench

Examples 5, 6, and 7 qualitatively compare the outputs of the base QwQ model and our QwQ-Med-3 model. The base model tends to think longer and often in repetitive loops. These examples also illustrate the reasoning-recall gap, where the base model is successful in recalling the entities along the KG path but is often unsure or unable to compose them into coherent reasoning, leading to incorrect conclusions. On the other hand, our curriculum-tuned models demonstrate the characteristic of structured step-by-step thinking traces derived from KG paths, and decisively reason with fewer tokens.

Example 5: QwQ-32B-Base vs. QwQ-Med-3 Response

KG Path from Diseases of the Digestive System:

Juvenile Colonic Polyposis $\xrightarrow{\text{maybe allelic with}}$ Hereditary haemorrhagic telangiectasia $\xrightarrow{\text{maybe causes}}$ Arteriovenous malformation $\xrightarrow{\text{maybe causes}}$ Intracranial arteriovenous malformation

Question:

A 14-year-old male presents with recurrent episodes of painless rectal bleeding and iron deficiency anemia. Colonoscopy reveals numerous hamartomatous polyps throughout the colon. Further investigation reveals scattered telangiectasias on his lips and nasal mucosa. Which of the following neurovascular abnormalities should be considered in this patient, given the constellation of findings?

Options:

A. Moyamoya disease B. Cerebral amyloid angiopathy **C. Intracranial arteriovenous malformation** D. Cavernous malformations

QwQ-Base-32B Response (1923 thinking tokens, incorrect answer):

`<think>` Okay, let's tackle this question. The patient is a 14-year-old male with painless rectal bleeding, iron deficiency anemia, and colonoscopy showing hamartomatous polyps throughout the colon. Additionally, there are telangiectasias on his lips and nasal mucosa. The question is asking which neurovascular abnormality to consider here.

First, I need to recall what conditions present with these symptoms. The key points here are the hamartomatous polyps in the colon and the telangiectasias. Hamartomatous polyps make me think of a hamartomatous polyposis syndrome. The most common ones are Peutz-Jeghers syndrome, Juvenile polyposis syndrome, and maybe Neurofibromatosis. But wait, the presence of telangiectasias might point to something else.

Peutz-Jeghers syndrome is characterized by hamartomatous polyps in the GI tract and mucocutaneous pigmentation, usually around the lips, mouth, and genital areas. The pigmentation is often described as "perioral" and "periungual." However, in this case, the telangiectasias are mentioned, which are small dilated blood vessels. Hmm, maybe I need to think of a different condition here.

Wait, another syndrome that includes both polyps and telangiectasias is Hereditary Hemorrhagic Telangiectasia (HHT), also known as Osler-Weber-Rendu disease. But HHT is primarily characterized by telangiectasias and arteriovenous malformations (AVMs) in various organs, including the lungs, liver, and brain. However, do HHT patients get hamartomatous polyps? I think HHT is more about telangiectasias and AVMs, but not necessarily polyps. So maybe that's not the right direction.

Wait, back to the polyps. Let me think again. The question mentions hamartomatous polyps. Juvenile polyps are usually not hereditary unless in Juvenile polyposis syndrome. But the key here is the combination with telangiectasias. Alternatively, there's a condition called PTEN Hamartoma Tumor Syndrome, which includes Cowden syndrome. Cowden syndrome can have multiple hamartomatous polyps and various other manifestations, including vascular malformations. But I'm not sure about telangiectasias specifically.

Alternatively, maybe the patient has a condition that includes both gastrointestinal polyps and neurological issues. Let me think of the possible neurovascular abnormalities listed in the options. The options are Moyamoya, cerebral amyloid angiopathy, intracranial AVM, or cavernous malformations.

Wait, the presence of telangiectasias and polyps might point to a genetic disorder that includes vascular abnormalities. Let me consider Turcot syndrome, but that's more about colorectal cancer and brain tumors like medulloblastoma or glioblastoma, so maybe not.

Wait, the telangiectasias on lips and nasal mucosa plus hamartomatous polyps. Hmm. Let me think again of Peutz-Jeghers. The pigmentation in Peutz-Jeghers is actually dark spots (due to increased melanin), not telangiectasias. So maybe the telangiectasias here are a clue towards another condition.

Alternatively, maybe the patient has a condition called Ataxia-Telangiectasia, but that's more about cerebellar ataxia, telangiectasias in eyes, and immunodeficiency. But the GI polyps aren't part of that.

Wait, the question mentions that the polyps are hamartomatous. Hamartomatous polyps in the colon could be part of a syndrome that includes vascular malformations. Let's think of Neurocutaneous melanosis? No, that's different.

Wait, the options given are neurovascular abnormalities. Let me look at the options again.

Option C is intracranial arteriovenous malformation. HHT is associated with AVMs, so if the patient had HHT, then C could be correct. But does HHT present with hamartomatous polyps? I think HHT does have telangiectasias and AVMs but not hamartomatous polyps.

Wait, maybe the key is that the hamartomatous polyps plus telangiectasias are part of a different syndrome. Let me think again.

Wait, there's a condition called PTEN Hamartoma Tumor Syndrome (Cowden syndrome) which can have gastrointestinal hamartomas, mucocutaneous lesions (like papillomatous papules, not sure about telangiectasias), and also can have neurological issues. For example, they might have vascular malformations. But the neurovascular abnormalities here could be cavernous malformations?

Wait, cavernous malformations are vascular lesions composed of blood-filled cavities separated by thin walls. They can occur in various conditions.

Alternatively, the Peutz-Jeghers syndrome doesn't have cavernous malformations.

Alternatively, the presence of telangiectasias and polyps might be pointing to a different genetic condition. Let me think of the genetic syndromes associated with both.

Wait, the options given include cavernous malformations (D). Let me think: some genetic conditions cause both GI polyps and cavernous malformations.

Alternatively, maybe the hamartomatous polyps and telangiectasias are part of a syndrome where cavernous malformations are common.

Alternatively, if the telangiectasias are part of HHT, then the neurovascular issue would be intracranial AVM (option C). But is that the case here?

Wait, the question says the patient has "scattered telangiectasias on his lips and nasal mucosa." HHT typically presents with mucocutaneous telangiectasias, especially on the lips, oral mucosa, and nose. So that fits. Also, HHT can have gastrointestinal telangiectasias leading to bleeding. But the question mentions hamartomatous polyps. Wait, do HHT patients get hamartomatous polyps? No, HHT is about telangiectasias and AVMs, not hamartomatous polyps. So that's conflicting.

Alternatively, maybe the telangiectasias are not due to HHT but another condition. The hamartomatous polyps are in the colon. Let me think again.

Another possibility is that the patient has a syndrome where both hamartomatous polyps and vascular malformations occur.

Wait, Cowden syndrome (PTEN mutation) can have hamartomatous intestinal polyps, as well as vascular malformations. In Cowden, they can have vascular malformations such as capillary malformations or cavernous malformations. So maybe cavernous malformations (option D) would be the answer here.

Alternatively, Neurofibromatosis type 1 can have GI polyps, but I don't recall them being hamartomatous in that context.

Wait, let me confirm: Cowden syndrome (PTEN-related) is associated with multiple hamartomatous polyps in the GI tract and also has an association with vascular malformations. Specifically, they can have cerebrovascular lesions like cavernous malformations. So the answer might be D, cavernous malformations.

Alternatively, another thought: the presence of telangiectasias and GI polyps might point to a different condition. Let me think of the possible neurovascular abnormalities. The options are Moyamoya (a specific vascular abnormality), cerebral amyloid angiopathy (usually in older patients), intracranial AVM (HHT), or cavernous malformations (Cowden).

Since the patient is 14, cerebral amyloid angiopathy is unlikely. Moyamoya is a progressive stenosis of the cerebral arteries, more common in children but not typically linked to these GI findings.

So between C and D. If the patient has HHT, then the AVM (C) would be the answer, but since the polyps are hamartomatous, which HHT doesn't have, then maybe the correct syndrome is Cowden, leading to cavernous malformations.

Alternatively, perhaps the telangiectasias are separate from the polyps, but the combination is pointing to Cowden.

Another angle: the question states "hamartomatous polyps throughout the colon" and "telangiectasias on lips and nasal mucosa". The combination of hamartomas and telangiectasias might be pointing to Cowden syndrome, which includes both. Since Cowden can have neurological issues including vascular malformations (like cavernous), then D is the answer.

Alternatively, if the telangiectasias are part of HHT, but the polyps are a coincidence? Unlikely, since the question is about the constellation of findings.

Therefore, the most likely answer is D, Cavernous malformations. `</think>` Final Answer: D

QwQ-Med-3 Response (820 thinking tokens, correct answer):

`<think>` Okay, let's break this case down. We have a 17-year-old female with a pretty complex history and current presentation that we need to piece together.

1. **Identify the Syndrome:** First, look at the patient's presentation. We have a 14-year-old male with recurrent painless rectal bleeding (often bright red blood per rectum, or 'painless hematochezia') and iron deficiency anemia. The colonoscopy confirms numerous hamartomatous polyps throughout the colon. Additionally, he has telangiectasias (dilated small blood vessels) on his lips and nasal mucosa. This specific combination of symptoms colonic hamartomatous polyps, rectal bleeding, and mucocutaneous telangiectasias is highly suggestive of **Hereditary Hemorrhagic Telangiectasia (HHT), also known as Osler-Weber-Rendu syndrome**.

2. **Understand HHT:** HHT is an autosomal dominant disorder affecting the vascular system. The classic features are:

   - **A:** Epistaxis (nosebleeds) often severe and recurrent.
   - **B:** Telangiectasias (on skin, lips, oral/nasal mucosa).
   - **C:** Gastrointestinal bleeding (from telangiectasias or sometimes larger arteriovenous malformations).
   - **D:** Pulmonary arteriovenous malformations (PAVMs).

   Other common features include liver AVMs and **cerebral vascular malformations**.

3. **Connect HHT to Neurovascular Abnormalities:** The question specifically asks about neurovascular abnormalities associated with this constellation of findings (which we've identified as likely HHT). HHT is known to cause various vascular malformations throughout the body, including the brain. The most common type of brain vascular malformation seen in HHT is an **intracranial arteriovenous malformation (AVM)**. These are abnormal connections between arteries and veins in the brain, bypassing the normal capillary system. They can cause symptoms like headaches, seizures, focal neurological deficits, or more seriously, intracranial hemorrhage (bleeding into the brain).

4. **Evaluate the Options:**

   - **A. Moyamoya disease:** This is a progressive narrowing (stenosis) and occlusion of the internal carotid arteries and their main branches, leading to the formation of a network of tiny collateral vessels (the "puff of smoke" or *moyamoya* appearance). While it causes strokes and TIAs, it's a distinct entity not directly linked to HHT.

- **B. Cerebral amyloid angiopathy (CAA):** This involves amyloid protein deposition in the walls of cerebral arteries, primarily affecting older adults (usually >60 years) and causing lobar hemorrhages. It's not associated with HHT.
- **C. Intracranial arteriovenous malformation (AVM):** This is a direct match for a known complication of HHT. The abnormal blood vessel formation characteristic of HHT frequently involves the brain, leading to AVMs. Given the patient's likely HHT diagnosis based on the colonic polyps and mucocutaneous telangiectasias, this is the most probable associated neurovascular abnormality.
- **D. Cavernous malformations:** These are collections of abnormal, thin-walled blood vessels. While they can occur in the brain and cause neurological symptoms or bleeding, they are not specifically characteristic of HHT like AVMs are.

5. **Conclusion:** Based on the patient's presentation (colonic hamartomatous polyps, mucocutaneous telangiectasias) which strongly points towards HHT, the most likely associated neurovascular abnormality to consider is an intracranial arteriovenous malformation (AVM). Therefore, option **C** is the correct answer.

`</think>` Final Answer: **C**

Example 6: QwQ-32B-Base vs. QwQ-Med-3 Response

KG Path from Mental, Behavioral, and Neurodevelopmental Disorders:

Fatal familial insomnia $\xrightarrow{\text{maybe cause}}$ Ataxia $\xleftarrow{\text{maybe cause}}$ MECR-related neurologic disorder $\xrightarrow{\text{maybe cause}}$ Choreoathetosis

Question:

A 55-year-old male presents with a rapidly progressive neurological decline over the past 6 months. His wife reports increasing insomnia, characterized by fragmented sleep with frequent, agitated awakenings. Neurological examination reveals dysarthria, myoclonus, and difficulty with gait. Brain MRI is unremarkable. Polysomnography demonstrates a significant reduction in slow-wave sleep and REM sleep. CSF analysis is normal except for slightly elevated levels of total protein. Genetic testing reveals a mutation in the PRNP gene. As his condition progresses, what additional neurological manifestation is MOST likely to develop given the underlying etiology?

Options:

A. Lower motor neuron weakness with fasciculations. B. Prominent resting tremor and rigidity. **C. Involuntary, writhing movements of the limbs and trunk.** D. Loss of vibratory sensation and proprioception in the lower extremities.

QwQ-Base-32B Response (7160 thinking tokens, incorrect answer):

`<think>` Okay, let's tackle this question step by step. The patient is a 55-year-old male with a 6-month history of rapidly progressive neurological decline. His symptoms include insomnia with fragmented sleep and agitation, dysarthria, myoclonus, and gait difficulty. The MRI of the brain is normal, which is a bit confusing because neurological issues often show up there, but maybe not in this case. The polysomnography shows reduced slow-wave and REM sleep. CSF is normal except for slightly high protein. Genetic testing found a PRNP mutation.

Hmm, PRNP gene mutations are associated with prion diseases, like Creutzfeldt-Jakob disease (CJD). I remember that prion diseases are transmissible spongiform encephalopathies. The key features here are the rapidly progressive dementia, myoclonus, and the sleep pattern changes. The sleep issues with reduced slow-wave and REM sleep fit with what I know about CJD. Also, MRI might not show changes early on, but sometimes in CJD, specific MRI findings like high signals in the cortex or basal ganglia can be seen, but the question says it's unremarkable. Maybe the MRI was done early in the disease course.

Looking at the options:

A. Lower motor neuron weakness with fasciculations. Lower motor neuron issues would suggest something like ALS, but prion diseases typically affect upper motor neurons more. ALS has both upper and lower, but the PRNP mutation points away from that.
B. Prominent resting tremor and rigidity. Resting tremor and rigidity are Parkinsonian symptoms. Parkinson's is another possibility, but the rapid progression and the genetic link to PRNP don't fit. Plus, CJD usually doesn't present with Parkinsonian features as a main symptom. The movement disorders in prion diseases are more like myoclonus or ataxia.
C. Involuntary, writhing movements. Writhing movements, like athetosis or chorea, are more typical of basal ganglia disorders, like Huntington's. But Huntington's has a different genetic mutation (HTT gene) and usually a slower progression. However, some prion diseases might have movement issues. Wait, but prion diseases can have ataxia, myoclonus, maybe some dystonia, but writhing movements? Not sure. Maybe not the most likely here.
D. Loss of vibratory sensation and proprioception. That sounds like sensory ataxia, which is seen in conditions like vitamin B12 deficiency, tabes dorsalis, or hereditary sensory neuropathies. In prion diseases, sensory deficits might occur, but more commonly, the cerebellar and cognitive symptoms are prominent. However, some forms of prion disease, like Gerstmann-Sträussler-Scheinker syndrome, can present with ataxia and maybe sensory issues. But the question mentions the PRNP gene mutation, which is more commonly associated with CJD. Also, the patient's initial presentation includes myoclonus and the sleep changes, which are classic for CJD. However, the question is asking what is "MOST likely to develop as the condition progresses".
Wait, in CJD, progression leads to more severe dementia, myoclonus, visual symptoms, maybe akinetic mutism. Sensory loss is less common. But let me think again. The options given: option D is about loss of vibratory and proprioception in lower extremities. That's more of a sensory neuropathy or dorsal column issue. In prion diseases, like in sporadic CJD, the main issues are cognitive, myoclonus, ataxia, visual problems. Maybe the sensory loss isn't a primary feature. Hmm, I'm going back and forth. Let me think of the options again. ....Given the PRNP mutation and the possibility of spinal cord involvement (even in rapidly progressive cases), the most likely additional manifestation is **D`</think>`** Final Answer: D

QwQ-Med-3 Response (1047 thinking tokens, correct answer):

`<think>` Okay, let's break this down.

1. **Identify the Diagnosis:** First off, we need to figure out what's going on with this patient. We have a 55-year-old male with a pretty rapid decline over just 6 months. The key features are the neurological problems like trouble speaking (dysarthria), jerky movements (myoclonus), and gait difficulty, plus the really specific sleep disturbance insomnia with lots of broken sleep and agitation. The brain MRI is clear, which is important because it helps rule out things like strokes or tumors. The CSF is mostly normal except for a tiny bit of protein, which isn't very specific. The real clincher here is the genetic test showing a mutation in the *PRNP* gene.

2. **What does the *PRNP* gene mutation mean?** This gene codes for the prion protein (PrP). Mutations in *PRNP* cause a group of diseases called prion diseases, which are fatal neurodegenerative disorders. The most common form you might know is Creutzfeldt-Jakob disease (CJD). However, different *PRNP* mutations can cause different clinical pictures, often referred to as familial prion diseases or familial CJD.

3. **Connecting Symptoms to Prion Disease:** The patient's symptoms fit well within the spectrum of prion diseases.
   - **Rapid Neurological Decline:** This is a hallmark of prion diseases.
   - **Myoclonus:** This is extremely common in prion diseases, often being a key feature.
   - **Dysarthria and Gait Difficulty:** These are also typical neurological signs.
   - **Sleep Disturbance:** The description of insomnia with fragmented, agitated sleep is characteristic of a specific type of prion disease called Fatal Familial Insomnia (FFI). FFI is caused by

certain specific mutations in the *PRNP* gene. The polysomnography results (reduced slow-wave and REM sleep) further support this type of sleep pathology seen in prion diseases.

- **Unremarkable MRI:** While some changes can eventually be seen, early in the course of many prion diseases, the MRI can be normal.

4. **Evaluating the Options:** The question asks what *additional* neurological manifestation is *most likely* to develop as his condition progresses. We already know he has myoclonus and dysarthria. Let's look at the choices in the context of prion disease progression:
    - **A. Lower motor neuron weakness with fasciculations:** This points towards motor neuron disease like ALS. While some rare overlap syndromes exist, isolated lower motor neuron signs aren't a typical primary feature of prion diseases.
    - **B. Prominent resting tremor and rigidity:** This sounds like Parkinsonism. Prion diseases can sometimes have Parkinsonian features (like slowness, stiffness), but a *prominent resting tremor* isn't characteristic.
    - **C. Involuntary, writhing movements of the limbs and trunk:** These are called **choreoathetoid movements**. This type of movement disorder is a known feature that can develop in various forms of prion disease as the illness progresses and affects different brain areas, particularly the basal ganglia. Given the underlying diagnosis of a prion disease (*PRNP* mutation), developing choreoathetosis is a plausible and relatively common additional neurological sign.
    - **D. Loss of vibratory sensation and proprioception in the lower extremities:** This points to sensory ataxia, often due to a problem in the posterior columns of the spinal cord. While prion diseases can affect sensory pathways (causing sensory loss) or cause spinocerebellar degeneration, this specific pattern isn't as classically associated with the progression of prion diseases as choreoathetosis is.
5. **Conclusion:** Based on the underlying diagnosis of a prion disease (*PRNP* mutation) and the known spectrum of neurological signs that can develop, involuntary writhing movements (choreoathetosis) are a well-recognized additional feature. Therefore, this is the most likely symptom among the choices to appear or become more prominent as the disease progresses.

`</think>` Final Answer: **C**

Example 7: QwQ-32B-Base vs. QwQ-Med-3 Response

KG Path from Neoplasms:

Adrenal Cortical Carcinoma $\xleftrightarrow{\text{may treat}}$ Mitotane $\xrightarrow{\text{maybe cause}}$ Hypercholesterolaemia

Question:

A 45-year-old female presents with virilization, new-onset diabetes mellitus, and a rapidly growing abdominal mass discovered on imaging. Hormonal evaluation reveals elevated cortisol, androgens, and estrogen levels. She is diagnosed with advanced stage adrenal cortical carcinoma and initiated on a standard treatment regimen. Six months later, follow-up labs show stable disease and controlled hormone levels. However, her lipid panel reveals a significant increase in total cholesterol and LDL cholesterol despite adhering to a low-fat diet. Which of the following is the MOST likely contributing factor to the patient's current hypercholesterolemia?

Options:

A. The direct effect of elevated cortisol on hepatic lipid metabolism. B. Prominent resting tremor and rigidity. **C. A side effect of the medication used to treat the adrenal cortical carcinoma.** D. Increased dietary intake of saturated fats secondary to improved appetite and overall well-being.

QwQ-32B-Base Response (3672 thinking tokens, incorrect answer):

`<think>` Okay, let's see. The question is about a 45-year-old female with adrenal cortical carcinoma presenting with virilization, new diabetes, and an abdominal mass. Her hormone levels are elevated in cortisol, androgens, and estrogens. After treatment, her disease is stable, but now she has high cholesterol and LDL despite a low-fat diet. The options are about why her cholesterol is up now.

First, I need to recall what adrenal cortical carcinoma does. These tumors can produce various hormones, like cortisol, androgens, estrogens, etc. Since her hormone levels were initially elevated, but now they're controlled (stable disease and hormone levels controlled), maybe the tumor isn't the direct cause now. But the question says the hypercholesterolemia appeared after treatment. Wait, the follow-up labs show stable disease and controlled hormones, so the tumor isn't actively secreting those hormones anymore? Or maybe they're controlled with treatment?

The cholesterol increase is despite a low-fat diet. So, the cause might be something else besides diet. Let's look at the options.

Option A: Elevated cortisol affecting hepatic lipid metabolism. But the question says her hormone levels are now controlled. If cortisol is under control, then maybe this isn't the cause now. Unless the treatment isn't fully controlling cortisol yet, but the question states "controlled hormone levels," so maybe A is not the answer.

Option B: Tumor-induced ectopic production of apolipoprotein B. ApoB is a key protein in LDL particles. If the tumor is producing ApoB, that could lead to more LDL. However, the tumor is now stable and maybe not as active. But the question says "tumor-induced ectopic" so maybe the tumor is producing this even when it's stable? Not sure. Alternatively, maybe the tumor's previous effects or other factors.

Option C: Side effect of medication. What's the standard treatment for adrenal cortical carcinoma? Let me think. Treatment might include mitotane, a drug used for adrenocortical carcinoma. Mitotane is an adrenal enzyme inhibitor, but does it affect cholesterol? Or maybe other meds like steroids? Wait, but she has hypercortisolism, so maybe they are using ketoconazole to suppress cortisol. Ketoconazole is an antifungal but also inhibits steroid synthesis. However, some treatments for cancer might include other drugs. Alternatively, if they're using glucocorticoids as part of treatment, but that might not be the case here. Alternatively, maybe statins? Wait, but she's on a low-fat diet, but the question is why her cholesterol is high. If the meds are causing it, like some immunosuppressants or other drugs, but I need to think of common meds for adrenal carcinoma that can raise cholesterol. Maybe mitotane? Not sure. Alternatively, maybe the treatment includes drugs that affect lipid metabolism. Hmm.

Option D: Increased dietary intake due to improved appetite. The question says she's adhering to a low-fat diet, so this is less likely. The answer probably isn't D.

Back to option B. Apolipoprotein B is made in the liver and is part of VLDL and LDL. If the tumor is producing ApoB ectopically, that could increase lipoproteins. But adrenal cortical carcinoma – do they produce ApoB? I'm not sure. Alternatively, maybe the tumor's androgen or estrogen production could influence lipid levels. But the hormone levels are now controlled, so that's not it. Wait, but maybe the tumor itself, even if stable, is producing something else?

Alternatively, maybe the standard treatment includes a drug that causes hyperlipidemia. Let's think: Mitotane can have side effects like fatigue, nausea, vomiting, but does it affect cholesterol? I don't recall. Let me think about other drugs. Chemotherapy agents might not directly cause high cholesterol. Alternatively, if the patient is on glucocorticoids, but the question states that her cortisol levels are elevated initially but now controlled, so maybe they are suppressing her own cortisol, hence not giving extra steroids. Alternatively, maybe the treatment includes ketoconazole which can have effects on cholesterol? Wait, ketoconazole inhibits steroid synthesis, but also might interfere with cholesterol synthesis? Not sure. Alternatively, maybe the hypercholesterolemia is a paraneoplastic effect even with stable disease. But the tumor is stable, so perhaps it's producing something else.

Alternatively, the tumor might be producing lipids or hormones that stimulate the liver to make more cholesterol. But the androgens and estrogens were controlled. Wait, adrenal cortical carcinoma can sometimes produce various steroids. If the tumor is still making some hormones, even if levels are controlled, maybe through treatment, but perhaps the treatment isn't fully suppressing all hormone production. Wait, the question says "controlled hormone levels," so probably not.

Alternatively, the tumor might be producing apolipoprotein B directly as an ectopic production. ....If I can't recall mitotane causing hypercholesterolemia, then B is the better option. Since the tumor could produce apolipoprotein B as an ectopic product, leading to increased LDL. Therefore, the most likely answer is B.
`</think>` Final Answer: B

QwQ-Med-3 Response (929 thinking tokens, correct answer):
`</think>` Okay, let's break down this question about the patient with adrenal cortical carcinoma (ACC) and hypercholesterolemia.

1. **Understand the Patient's Situation:** We have a 45-year-old female diagnosed with advanced ACC. This is a rare cancer of the adrenal cortex. Key things to note are her initial presentation (virilization, new-onset diabetes, high cortisol, androgens, estrogens) and the treatment course. She was treated, and her disease is stable, and importantly, her hormone levels (cortisol, androgens, estrogens) are now controlled. However, her lipid panel shows high cholesterol and LDL cholesterol, even though she's on a low-fat diet. We need to figure out the most likely reason *why* her cholesterol is high *now*.

2. **Analyze the Options:**
   - **A. The direct effect of elevated cortisol on hepatic lipid metabolism:** This is a very plausible mechanism. Chronic elevation of cortisol (a glucocorticoid) is well-known to mess with liver metabolism. Cortisol increases the production of Very Low-Density Lipoproteins (VLDL) by the liver. VLDL is rich in triglycerides, but as it travels through the blood, its triglycerides are removed, and it gets converted into LDL (the "bad" cholesterol). So, more VLDL means higher LDL. Furthermore, cortisol can also decrease the number of LDL receptors on cells, meaning the body can't clear LDL from the blood efficiently. This is often seen in conditions like Cushing's syndrome, which is characterized by excess cortisol. *However*, the case explicitly states her hormone levels are now *controlled*. While controlling a severe hypercortisolism might take time or might not fully normalize metabolic effects immediately, we *can't* assume her cortisol is still significantly elevated enough to be the *primary* driver of this new hypercholesterolemia, especially since she was treated and labs show controlled levels.
   - **B. Tumor-induced ectopic production of apolipoprotein B:** Apolipoprotein B (ApoB) is a key protein component of VLDL and LDL. If the tumor itself were somehow producing ApoB ectopically (in an abnormal place), it could theoretically lead to increased VLDL/LDL. However, this is a highly specific and unusual mechanism. While some tumors can cause paraneoplastic syndromes (abnormal conditions caused by the tumor's presence, not its direct invasion), significant ectopic ApoB production leading to hypercholesterolemia isn't a recognized or common feature of ACC.
   - **C. A side effect of the medication used to treat the adrenal cortical carcinoma:** This is a very strong possibility. Advanced ACC is often treated with chemotherapy, and a common first-line agent is Mitotane (o,p'-DDD). Mitotane is known to have various side effects, including metabolic ones. Importantly, hypercholesterolemia (increased cholesterol and LDL) is listed as a potential adverse effect of Mitotane. Since the patient was diagnosed, started treatment (likely including Mitotane), and developed hypercholesterolemia *after* the treatment initiation (and her other hormone levels became controlled), a medication side effect becomes a highly likely explanation. The timing fits perfectly.
   - **D. Increased dietary intake of saturated fats secondary to improved appetite and overall well-being:** While it's true that dietary saturated fats contribute to high cholesterol, the patient is explicitly stated to be adhering to a *low-fat diet*. This makes dietary intake the least likely culprit for a *significant* increase in her lipids.

3. **Synthesize and Conclude:** We've ruled out dietary intake (D) because of the low-fat diet. Ectopic ApoB production (B) is too rare and specific. Elevated cortisol (A) could cause it, but

her hormones are stated to be controlled. This leaves the medication side effect (C) as the most probable cause. Given that advanced ACC is often treated with Mitotane, and Mitotane can cause hypercholesterolemia, this aligns best with the clinical scenario presented, especially considering the timeline (treatment initiated, then hypercholesterolemia developed despite controlled hormones).

Therefore, the MOST likely contributing factor is a side effect of the medication used to treat the adrenal cortical carcinoma.

`</think>` Final Answer: **C**

E Additional Details: ICD-Bench Results across Task Difficulty

We estimated task difficulty using the *pass*@1 score of the QwQ-32B base model over 16 independent samples. The score distribution across the $3,675$ ICD-Bench questions is shown in Fig. 13 (top). We observed a large proportion of full scores, followed by a heavy tail at a zero pass rate. We empirically segmented the questions into five difficulty bins, such that the resulting accuracy of the base model decreases roughly linearly across the bins. In addition, we also quantified the average KG path length of the question for each score in Fig. 13 (bottom) and found no meaningful correlation between the path lengths and the ability of the base model to solve the question. In Fig. 14, we present the ICD-Bench results divided across difficulty levels within each category. We observe that the overall decline in performance with increasing difficulty is consistent across most categories and all models, with curriculum-tuned models exhibiting a widening performance gap at the harder levels.

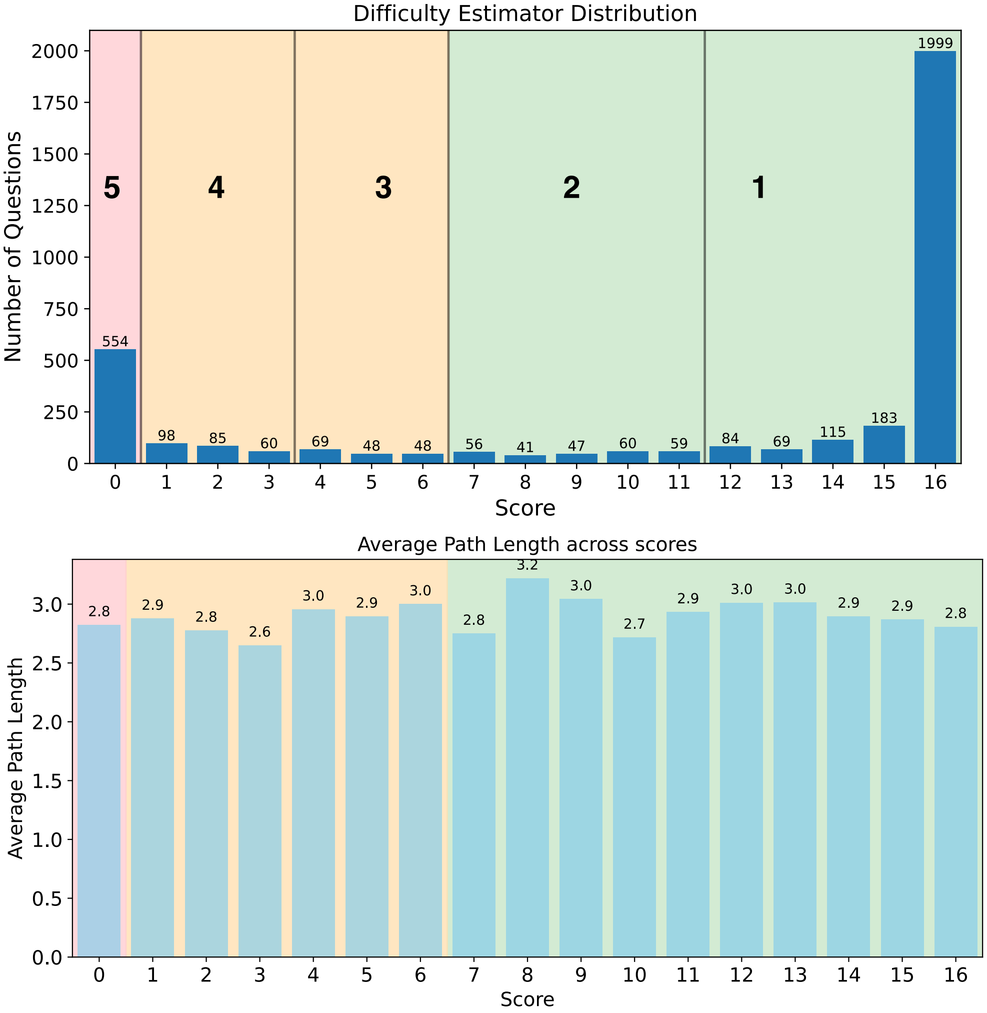


Figure 13: Top: Distribution of *pass*@1 score of the QwQ-Base-32B model over 16 samples, across ICD-Bench questions. Bottom: Mean KG-path length of questions across *pass*@1 scores.

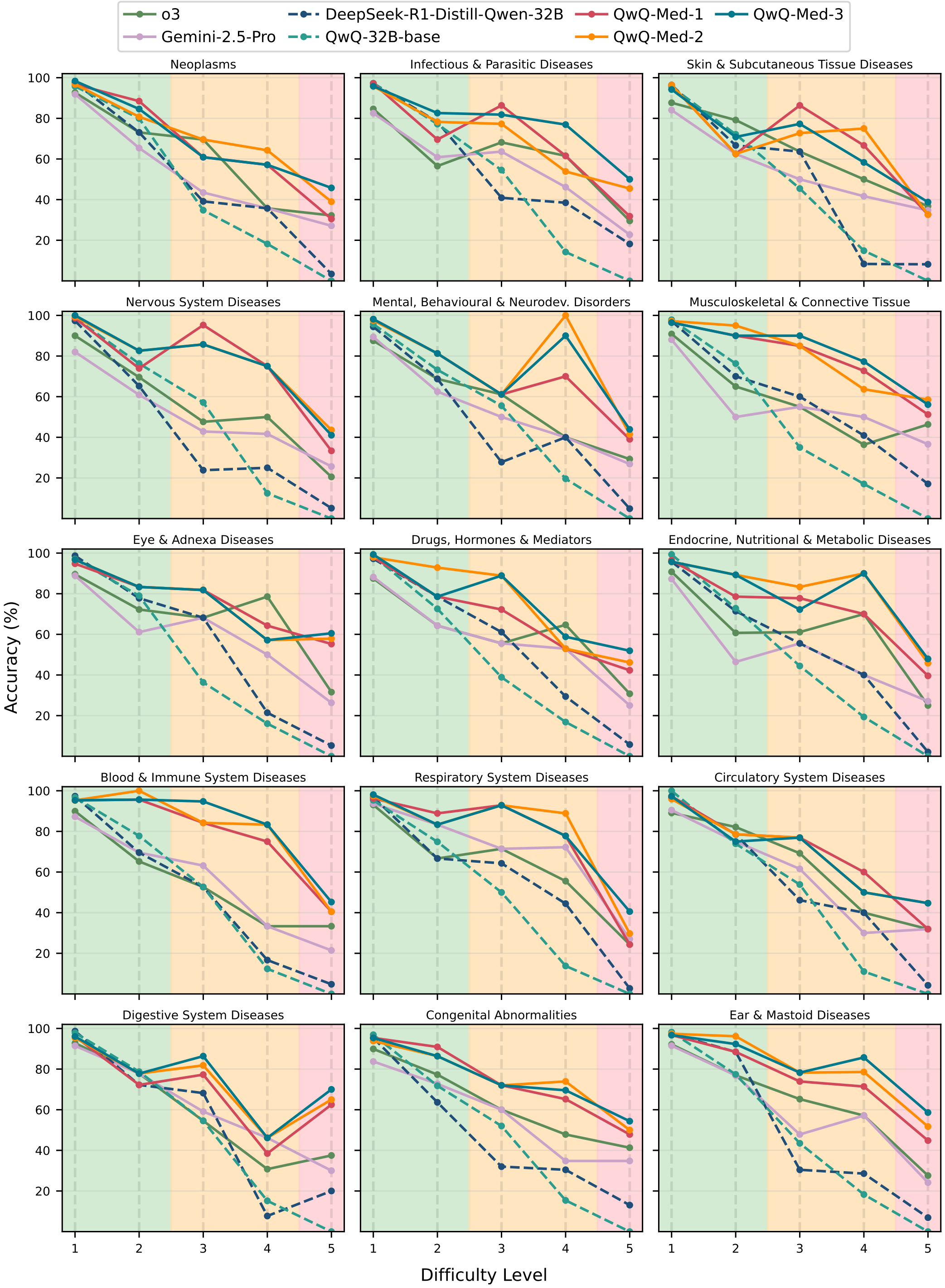
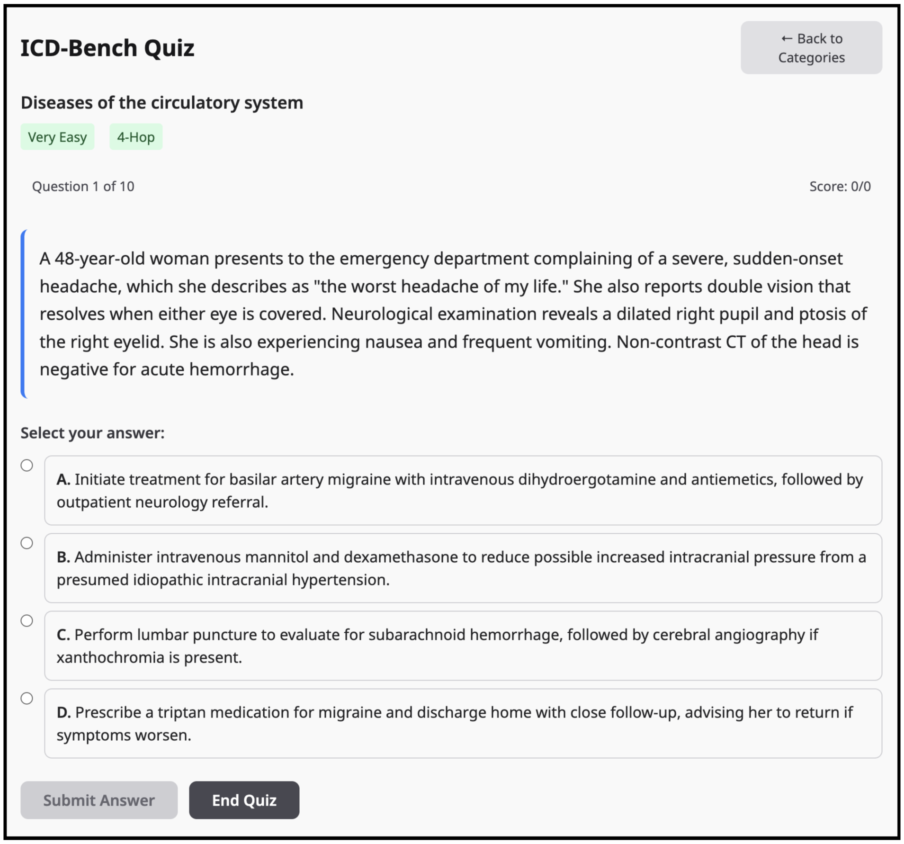


Figure 14: ICD-Bench performance of different models across difficulty levels within each ICD-category.

F Human Evaluations

We evaluated the performance of humans on ICD-Bench to estimate the relative expertise of our curriculum-tuned models. To do so, we designed a quiz administered to a cohort of humans via a web interface. Users were given the option to select a domain for the quiz and were administered 10 questions at random across the five difficulty levels for that domain. Fig. 15 shows a sample question as it appears on our web interface.



A 48-year-old woman presents to the emergency department complaining of a severe, sudden-onset headache, which she describes as "the worst headache of my life." She also reports double vision that resolves when either eye is covered. Neurological examination reveals a dilated right pupil and ptosis of the right eyelid. She is also experiencing nausea and frequent vomiting. Non-contrast CT of the head is negative for acute hemorrhage.

○ **A.** Initiate treatment for basilar artery migraine with intravenous dihydroergotamine and antiemetics, followed by outpatient neurology referral.

○ **B.** Administer intravenous mannitol and dexamethasone to reduce possible increased intracranial pressure from a presumed idiopathic intracranial hypertension.

○ **C.** Perform lumbar puncture to evaluate for subarachnoid hemorrhage, followed by cerebral angiography if xanthochromia is present.

○ **D.** Prescribe a triptan medication for migraine and discharge home with close follow-up, advising her to return if symptoms worsen.

Figure 15: A web interface to administer ICD-Bench questions to humans via a quiz: a sample question as it appears on the interface, showing the question, difficulty, hop length, category, and quiz score.

We collected 924 responses across eight ICD categories from a cohort of medical experts, including medical students and residents, as well as practicing clinicians. In Fig. 16, we compare human performance against that of QwQ-Med-3, broken down across these ICD categories. Similarly, in Fig. 17, we compare QwQ-Med-3 against human performance across difficulty levels. Our major findings are:

- **QwQ-Med-3 consistently outperforms humans across nearly all ICD-10 categories, showing higher and more consistent performance**. This suggests that the model surpasses human diagnostic ability across diverse medical domains. Its performance advantage is especially notable in categories like endocrine diseases, drugs and mediators, and infectious diseases, indicating strong domain coverage and reliability for these categories.

- **Categories difficult for humans are not necessarily difficult for the model.** QwQ-Med-3 maintains high accuracy even in fields where the performance of our human cohort drops, implying that model errors are not aligned with human weaknesses. This divergence highlights the potential of the model to complement human expertise in clinical reasoning.

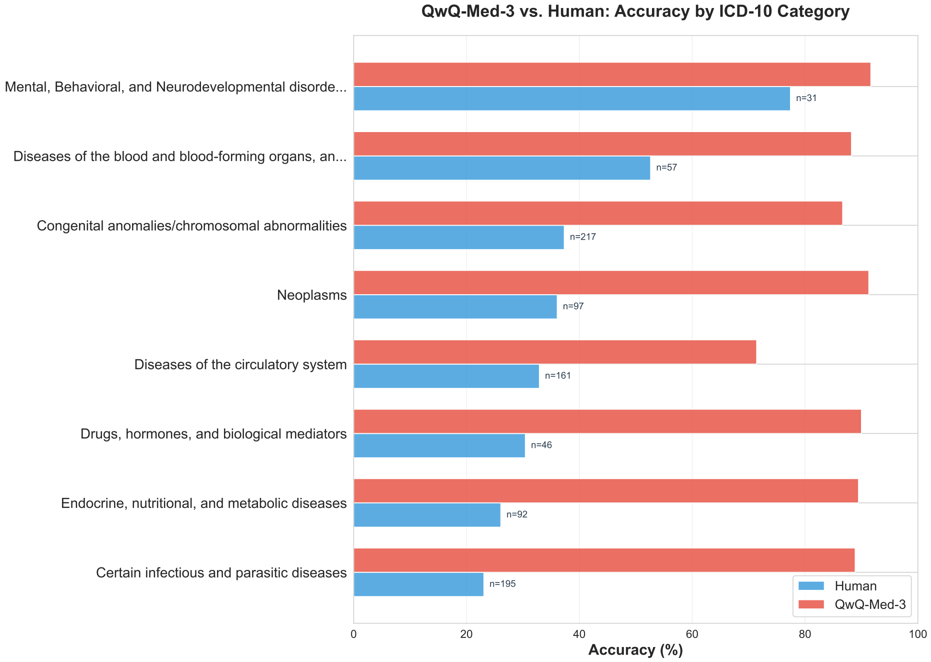


Figure 16: Breakdown of human performance against QwQ-Med-3 across ICD-10 categories. We only include categories where we received at least 30 human responses. 'n' here refers to the number of responses received from the cohort of humans for the corresponding ICD-10 category.

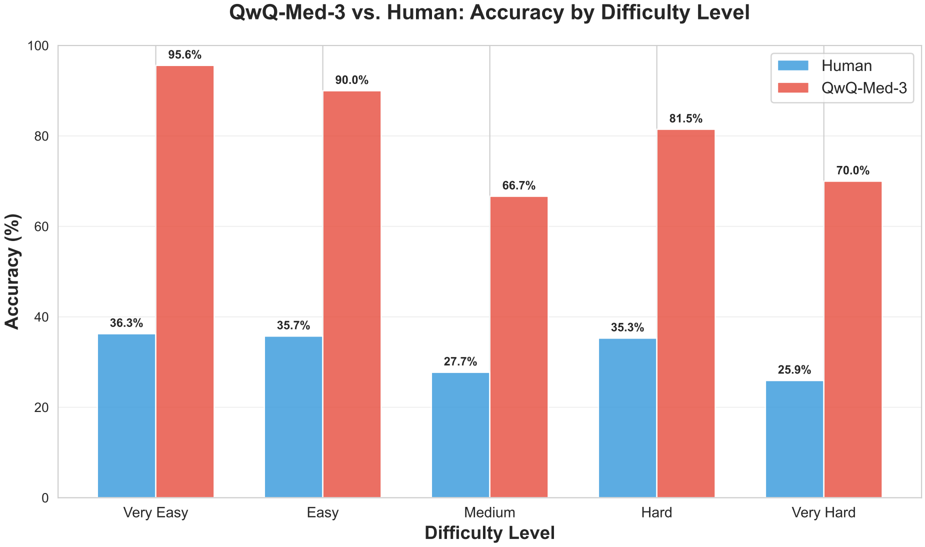


Figure 17: Breakdown of human performance against QwQ-Med-3 across ICD-Bench difficulty levels.

- **The difficulty level impacts both humans and the model similarly, but QwQ-Med-3 performs better even on harder questions.** As can be seen from Fig. 17, accuracy typically declines as the difficulty level increases for both the QwQ-Med-3 model and the human cohort. Yet, the relative performance advantage of the model stays robust at harder difficulty levels. This suggests that QwQ-Med-3 handles complex reasoning and ambiguous clinical cues more effectively than the human participants.

G Additional Details: Medical QA Benchmarks

Prompt 4 shows the template used to generate responses from both state-of-the-art reasoning and non-reasoning models on existing medical benchmarks. Table 3 lists the hyperparameters employed during model inference. Expanded performance results on various MMLU-Med subsets are presented in Table 4, including available results on proprietary models, as reported in (Yun et al., 2025).

Prompt 4: Template for Medical QA Benchmarks

```
You are a medical expert presented with an MCQ question. Your final answer should (A, B, C, or D)
  <Question>
      [Clinical Vignette]
  </Question>
  <Options>
      A. [Option]
      B. [Option]
      C. [Option]
      D. [Option]
  </Options>
```

Model	MAX_NEW_TOKENS	MAX_LENGTH	TEMPERATURE	`TORCH_DTYPE`	REPETITION_PENALTY
MedGemma	2048	8192	0.6	bfloat16	—-
Meerkat	—-	8192	0.7	bfloat16	1.2
Deepseek	—-	8192	0.6	bfloat16	—-
Qwen3	—-	8192	0.6	bfloat16	1.2
SkyT1	—-	8192	0.6	bfloat16	—-
QwQ	—-	8192	0.6	bfloat16	—-

Table 3: Hyperparameters used for model benchmarking. Unless specified otherwise, default hyperparameters from model providers were used. We used Hugging Face .generate() defaults.

Model	MedQA USMLE	PubMed QA	Med MCQA	MMLU (Med-Avg)	MMLU Clinical	MMLU Genetics	MMLU Anatomy	MMLU Prof Med	MMLU College Med	MMLU College Bio
					Proprietary Models (API only)					
Gemini Flash 2.0	87.51	–	72.60	92.01	–	–	–	–	–	–
GPT-4o-Mini	79.03	–	68.20	87.79	–	–	–	–	–	–
o4-mini	93.95	–	79.60	93.99	–	–	–	–	–	–
o3-mini	92.69	–	75.50	93.01	–	–	–	–	–	–
					Open-Source Models					
R1-Distill-Qwen (32B)	74.31	76.00	62.23	85.08	85.61	88.89	78.36	87.08	79.65	90.91
Qwen3 (32B)	64.34	65.40	55.63	69.53	73.48	72.73	57.46	69.37	65.12	79.02
Sky-T1 (32B)	70.78	67.40	63.52	84.91	84.47	88.89	78.36	88.19	77.91	91.61
MedGemma (27B)	60.49	70.40	63.42	78.46	85.23	86.87	70.90	67.90	73.84	86.01
Meerkat (70B)	78.95	77.40	68.42	83.74	79.92	83.84	79.85	89.67	76.16	93.01
QwQ (32B)	85.62	71.00	69.26	90.46	88.26	98.00	82.09	90.77	85.47	97.20
QwQ/parallel-scal.	**87.09**	78.02	71.62	91.33	89.36	98.40	83.10	93.34	85.97	97.85
					Our Curriculum-Tuned Model					
QwQ-Med-3 (32B)	82.72	76.00	71.03	90.64	89.59	98.22	81.02	92.02	85.07	96.88
QwQ-Med-3/parallel-scal.	85.39	**78.19**	**73.25**	**92.90**	**91.10**	**98.88**	**84.84**	**95.09**	**89.25**	**98.01**

Table 4: Expanded results from evaluating state-of-the-art models on medical QA benchmarks. Our model, QwQ-Med-3, consistently achieves competitive performance across all benchmarks when compared to similarly sized open-source models. However, it lags behind larger proprietary models, likely due to differences in scale, training compute, and access to private data resources. Best performance on each benchmark is indicated in **bold**, while second-best is underlined, excluding proprietary models.

H Additional Related Work

Medical Question Answering. The testbed used for our ICD-Bench is medical question answering, a challenging task that requires models to exhibit reasoning and medical expertise. Traditional systems often rely on access to high-quality external knowledge bases and exhibit limited performance improvements due to the complexity of mapping natural questions to exact answers. The advent of LLMs has led to rapid progress in the field; researchers have introduced benchmarks like MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), MMLU (Hendrycks et al., 2020), and PubMedQA (Jin et al., 2019), derived from medical board exams and clinical topics, to evaluate LLMs on medical reasoning and factual recall. Multiple studies have demonstrated impressive capabilities of LLMs on USMLE and other medical exams, highlighting the benefits of scale and diverse pre-training. Nori et al. (2023) show GPT-4 exceeding the passing score by more than 20 points on USMLE and outperforming smaller, but fine-tuned, models. Google's Med-PaLM 2 (Google, 2024) and the MedGemma series (Google, 2025), specifically fine-tuned on medical data, achieve state-of-the-art performance on benchmarks like MedQA and MedMCQA. Kim et al. (2025), with its Meerkat family of models, exhibit the significant potential of small LMs to improve reasoning on medical tasks through careful data curation from verified sources.

The above results underscore how techniques like knowledge extraction, targeted instruction tuning, and inference-time reasoning collectively contribute to more robust medical QA and diagnostic systems. Yet, challenges regarding integrity and reliability persist; even top models hallucinate and sometimes offer incorrect or unsafe advice. Future work must integrate knowledge-grounded reasoning with rigorous validation to advance domain-specific capabilities while exhibiting the trustworthiness required in critical settings. To address this need, we present a new paradigm for grounding LLMs with verified domain-specific knowledge through KGs and LoRA fine-tuning, thereby enhancing their multi-hop reasoning capabilities.

Inference-time Scaling. Beyond the pre- and post-training, inference-time scaling techniques aim to boost reasoning performance of LLMs during inference. Frontier thinking models, such as OpenAI's o1 and Gemini 2.5 Pro, are known to employ this technique to achieve better results. This involves allocating additional compute or steps at query time to elicit deeper reasoning (Muennighoff et al., 2025). One paradigm is to prompt the model to produce explicit step-by-step solutions, known as chain-of-thought prompting (Wei et al., 2023; Wu et al., 2024). More advanced inference-time strategies treat the LLM as an agent that iteratively refines or verifies its own answers (Google Deepmind, 2024). So-called "external" inference-time scaling methods employ additional models or search procedures alongside the main LLM. Broadly, these techniques can be classified into sequential or parallel inference-time strategies. For example, a best-of-N strategy simply generates N candidate answers in parallel with reasoning and picks the most promising one using a scoring function or a separate verifier model. Other approaches perform a sequential beam search over reasoning steps or use a tree-of-thoughts where the model explores multiple branches of reasoning and a voting or value function identifies the best path (Bi et al., 2024; Franceschelli & Musolesi, 2024).

A recent study by Liu et al. demonstrates that with the right inference-time scaling, a relatively small 1B-parameter model can outperform a 405B-parameter model on complex math problems (Liu et al., 2025). Scaling inference-time compute offers a promising direction to improve the capability of LLMs on complex multi-hop reasoning tasks. We employ both parallel and sequential inference-time scaling strategies with our model, demonstrating improvement across ICD-Bench.

Synthetic Data Generation from Structured Sources. Structured data often imbue crucial domain-specific priors through the specific structural organization of domain data. This domain structure can be further used to synthetically generate and augment training data that capture diverse primitives, as demonstrated in this work based on a KG. Past works explicitly elicit compositional and zero-shot generalization with large KGs (Gu et al., 2021; Dubey et al., 2019), scale synthetic augmentation (Dubey et al., 2019), and create holistic evaluations (Tatarinov et al., 2025). Beyond KGs, researchers have used tabular SQL schemas (Deng et al., 2021; Yu et al., 2021; Zhong et al., 2017) and lists (Amouyal et al., 2023; Malaviya et al., 2023) to source QA pairs across domains. Beyond a predefined structure, researchers have also attempted to induce structure in datasets by linking thematically-related documents and text chunks (Welbl et al., 2018), combining bottom-up curriculum with top-down natural language aggregation.

The choice of structured data source should align closely with the nature of the downstream task and the structure of the target domain. For instance, when the data are relatively flat, with many categorical attributes, tabular SQL schemas provide a natural fit, as columns can capture discrete category relationships and constraints across entities. KGs, on the other hand, are more suitable for domains that have hierarchical ontologies or richly relational data, where entities and their interdependencies play a central role. Lists and enumerations are particularly effective in representing ordered or set-like information, such as event sequences, itemized facts, or aggregated entity groups.