

Towards Efficient Medical Reasoning with Minimal Fine-Tuning Data

Supplementary Material

Appendix Contents

A. Details of Pilot Experiment	1
B. Case Study	2
C. Details of Difficulty Score	2
C.1. Prompt for Difficulty Score Annotation	2
C.2. Difficulty Classifier Training	4
C.3. Difficulty Score Distribution	4
D. Details of Influence Score	4
D.1. Influence Score Computation	4
D.2. Influence Score Distribution	4
E. Details of Clinical Value Assessment	4
F. Details of Efficiency Analysis	8
G. Full Experimental Results	10
G.1. Main Results	10
G.2. QA Experiment	10
G.3. Ablation Study Experiment	11
G.4. Generalization Experiment	11

A. Details of Pilot Experiment

To investigate the interplay between medical difficulty and sample influence, we conducted a pilot experiment on the FineMed dataset. We first partitioned the data into four quadrants (We set difficulty score threshold as 3 and choose the median score of influence scores as bounder) based on difficulty and influence scores: Q_1 with high difficulty and high influence, Q_2 with low difficulty and high influence, Q_3 with high difficulty and low influence, and Q_4 with low difficulty and low influence. Subsequently, we fine-tuned separate instances of the Qwen3-8B model, each using a 1% data subset drawn exclusively from one of the four quadrants. The impact of each quadrant was assessed through a two-pronged evaluation: 1) qualitative reasoning ability, scored on a 5-level Likert scale by Gemini-2.5-pro, and 2) quantitative task performance, measured by accuracy across nine downstream datasets. The full results of pilot experiment are shown in Table 5. The prompt for the reasoning quality evaluation is provided below.

Reasoning Quality

You are an experienced medical doctor and your task is to systematically evaluate and score the clinical reasoning process.

I. Aspects to Consider for Evaluation

When reading and analyzing a medical reasoning text, please consider the following three core areas

holistically:

1. Analysis and Reasoning Process

Completeness of Information: Was all key clinical information (history, signs, lab and imaging results, etc.) accurately and comprehensively identified?

Synthesis of Information: Was scattered data (symptoms, risk factors, test results) effectively synthesized into a coherent and meaningful clinical picture?

Logical Chain: Is the reasoning process clear, rigorous, and progressive? Are there any logical leaps or contradictions?

Differential Diagnosis: Were other relevant possibilities (key differential diagnoses) considered and reasonably ruled out based on the available evidence?

2. Application of Knowledge

Accuracy of Knowledge: Is the applied medical knowledge (e.g., pathophysiology, epidemiology, drug mechanisms) accurate?

Adherence to Guidelines: Does the understanding of diagnostic criteria and treatment options align with current, accepted clinical guidelines and evidence-based medicine?

3. Conclusion and Justification

Correctness of Conclusion: Is the final diagnosis and proposed management plan correct?

Quality of Justification: Is the reasoning provided for the final conclusion clear, persuasive, and well-supported by the evidence in the case?

II. Comprehensive Scoring Rubric (1-5 Points)

After holistically considering all the points above, assign a single comprehensive score that best reflects the overall quality, based on the following criteria:

5 (Excellent): The reasoning process is exemplary. The analysis is thorough, the logic is flawless, the application of knowledge is precise, and the conclusion is correct and exceptionally well-justified. It mirrors the thinking of an expert clinician.

4 (Good): The reasoning process is strong and leads to the correct conclusion. The core logic and knowledge are sound, but there may be minor omissions in how the process is presented (e.g., not fully elaborating on the differential diagnosis), without affecting the overall outcome.

3 (Adequate): The reasoning arrives at the correct conclusion, but the process has noticeable shortcomings. The logical chain may be unclear, the justification weak, or it may rely more on "pattern matching" than systematic analysis. It answers "what" but not "how" or "why."

2 (Poor): The reasoning process has significant flaws. It may miss key data, apply incorrect knowledge, or follow a convoluted logical path, often leading to an incorrect or incomplete conclusion.

1 (Very Poor): The reasoning is fundamentally flawed, demonstrating a lack of basic understanding of the clinical scenario, significant knowledge errors, and a complete absence of logical structure. The conclusion is unsubstantiated.

Please use the following format for your response.

- Score: [1-5]
- Rationale: [Provide a brief, specific justification for the score, citing examples from the response.]

Here are the Question and Answer:

B. Case Study

As shown in Figure 7, we provide a case study of Qwen3-8B trained on DIQ-1% FineMed answering a question in MedBullets-option5 and mark the parts of *Differential Diagnosis (DDx)*, *Safety Check*, and *Evidence Citation* in **Red**, **Orange**, and **Blue**.

The model employs a systematic and evidence-based approach to clinical problem-solving. It initiates its analysis by correlating the patient's history and risk factors with key laboratory findings, principally the profoundly low CD4⁺ count, to establish a diagnosis of severe immunosuppression. This correctly frames the presenting problem within the context of an opportunistic central nervous system infection. Subsequently, the model focuses on the most diagnostically salient evidence from the lumbar puncture. It interprets the cerebrospinal fluid (CSF) profile—characterized by lymphocytic pleocytosis, hypoglycorrachia (low glucose), and elevated protein—as highly suggestive of a fungal etiology. The positive India ink stain is correctly identified as the definitive finding that confirms a diagnosis of cryptococcal meningitis. Finally, in determining the management plan, the model assesses the disease's severity. It logically selects the standard-of-care induction therapy for severe cryptococcosis, Amphotericin B and flucytosine, while correctly distinguishing this from treatments for other pathogens or from therapies, such as fluconazole, which are reserved for less severe presentations or consolidation phases.

C. Details of Difficulty Score

C.1. Prompt for Difficulty Score Annotation

We provide the prompt for obtaining medical difficulty scores along three dimensions (*Knowledge*, *Reasoning*, and *Overall*) in the following box.

Medical Difficulty

You are an experienced medical doctor and independent practitioner. Your task is to classify a medical question across THREE dimensions following a specific evaluation sequence: First assess Knowledge Complexity, then Reasoning Complexity, and finally provide an Overall Difficulty rating that synthesizes both dimensions.

Evaluation Sequence: Knowledge → Reasoning → Overall

Please evaluate each dimension independently in the specified order, as this sequence ensures a more systematic and comprehensive assessment.

—

Dimension 1: Knowledge Complexity (1-5 Levels)

Classify based on the depth and breadth of medical knowledge required:

Level 1 (Basic Medical Knowledge): The question requires fundamental medical concepts taught in early medical education. Common diseases, basic anatomy/physiology, standard definitions.

Level 2 (Standard Clinical Knowledge): The question requires typical clinical knowledge expected of practicing physicians. Common clinical presentations, standard diagnostic criteria, routine management principles.

Level 3 (Specialty Foundational Knowledge): The question requires specialized knowledge within specific medical fields. Subspecialty concepts, advanced pathophysiology, specialized diagnostic approaches.

Level 4 (Deep Specialty Knowledge): The question requires expert-level knowledge within specialized domains. Rare diseases, complex pathophysiology, advanced subspecialty management, cutting-edge diagnostic techniques.

Level 5 (Cutting-edge/Rare Specialized Knowledge): The question requires knowledge of very rare conditions, latest research findings, experimental treatments, or highly specialized expert-level concepts that even specialists might need to reference.

—

Model	MedQ	MedM	MMLU	Avg _S	HLE	MeB4	MeB5	MedX	MedG	MetM	Avg _C	Avg _A	Reasoning Quality
1% Q ₁	78.01	64.89	83.93	75.61	15.53	66.56	59.09	18.69	55.07	62.78	46.29	56.06	4.82
1% Q ₂	76.04	56.30	83.93	<u>72.09</u>	9.71	64.94	58.12	17.88	54.59	59.72	<u>44.16</u>	<u>53.47</u>	4.27
1% Q ₃	66.85	65.26	83.10	71.74	9.71	62.34	57.79	14.53	54.27	58.99	42.94	52.54	<u>4.60</u>
1% Q ₄	60.53	53.79	74.10	62.81	12.62	62.23	58.44	12.78	48.24	58.49	42.13	49.02	4.18

Table 5. Full downstream task accuracy and reasoning quality results of pilot experiment.

Dimension 2: Reasoning Complexity (1-5 Levels)

Classify based on the level of medical reasoning difficulty required:

Level 1 (Direct Recall/Understanding): The question primarily tests direct recall of medical facts, definitions, common associations, or basic recognition. It requires no complex reasoning; the answer is a straightforward retrieval of memorized knowledge.

Level 2 (Simple Application): The question requires basic application of well-established medical knowledge to straightforward scenarios. Involves simple pattern recognition or direct application of standard protocols with minimal reasoning steps.

Level 3 (Moderate Reasoning): The question requires applying medical knowledge to specific, often slightly novel, scenarios. It involves interpreting clinical data, making logical connections between symptoms and conditions, or performing straightforward differential diagnosis. It typically involves 2-3 clear reasoning steps.

Level 4 (Complex Reasoning): The question demands integration of multiple pieces of information from various domains (e.g., history, physical, labs, imaging), complex differential diagnosis, evaluation of multiple management options, or navigating moderately ambiguous data. It involves multi-step logical chains and synthesis of information.

Level 5 (Expert-level Reasoning/Complex Problem Solving): The question requires advanced clinical reasoning with high-level integration of complex, ambiguous, or incomplete data from multiple domains. It involves sophisticated differential diagnosis, evaluation of competing hypotheses, critical evaluation of conflicting information, and navigation of highly nuanced clinical scenarios. Requires expert-level clinical judgment and complex multi-step reasoning chains.

When determining reasoning level, consider:

- The amount of information provided in the ques-

tion (how many data points need integration)

- The number and complexity of reasoning steps required
- The degree of ambiguity or nuance present in the scenario
- Whether the answer derives from direct recall versus requiring deductive/inductive reasoning
- The sophistication of clinical judgment required

Dimension 3: Overall Difficulty (1-5 Levels)

Comprehensive assessment that synthesizes both Knowledge and Reasoning complexity:

Level 1 (Very Easy): Low knowledge requirements with minimal reasoning demands. Straightforward questions with clear answers, minimal clinical complexity, common scenarios.

Level 2 (Easy): Moderate knowledge requirements or simple reasoning, but not both simultaneously. Slightly more complex but still manageable scenarios.

Level 3 (Moderate): Balanced combination of knowledge and reasoning demands, or high complexity in one dimension compensated by lower complexity in the other. Moderate clinical complexity requiring integrated thinking.

Level 4 (Hard): High demands in both knowledge and reasoning, or extreme complexity in one dimension. Complex scenarios requiring advanced clinical judgment, significant ambiguity, multiple competing factors.

Level 5 (Very Hard): Exceptional demands in both knowledge and reasoning simultaneously. Extremely challenging scenarios requiring expert-level judgment, high ambiguity, multiple complex factors, potentially controversial or cutting-edge topics.

Overall Difficulty Synthesis Guidelines:

- Consider how Knowledge and Reasoning complexity interact
- High knowledge + high reasoning = very challenging
- High knowledge + low reasoning = moderate

- challenge
- Low knowledge + high reasoning = moderate challenge
- Account for cumulative cognitive load

Output Format:

Please provide your assessment in the following format:

Knowledge Complexity Score: [1-5]

Reasoning Complexity Score: [1-5]

Overall Difficulty Score: [1-5]

Knowledge Justification: [Explain the knowledge requirements - medical domain depth, specialization level, rarity of concepts, specific medical knowledge needed]

Reasoning Justification: [Explain the reasoning demands - information integration, logical steps, ambiguity handling, clinical reasoning complexity]

Overall Difficulty Justification: [Explain how Knowledge and Reasoning complexity combine to create the overall challenge level, considering their interaction and cumulative impact]

Key Factors:

- Primary difficulty drivers
- Interaction between knowledge and reasoning demands
- Clinical context considerations
- Any notable complexities or special considerations

Please evaluate the following medical reasoning question and note that you only need to evaluate the difficulty and you don't need to answer the question.

C.2. Difficulty Classifier Training

We evaluated three lightweight BERT-style models for predicting medical difficulty: BiomedBERT [6], ClinicalBERT [9], and ModernBERT [28]. As shown in Table 6, BiomedBERT consistently outperformed the other models and was therefore selected as our difficulty classifier.

Difficulty	BiomedBERT	ClinicalBERT	ModernBERT
Knowledge	80.89	76.91	78.69
Reasoning	83.86	82.69	83.55
Overall	81.90	80.84	81.05

Table 6. Test set F1 scores on difficulty classification task of three BERT-style models.

Influence	BiomedBERT	ClinicalBERT	ModernBERT
Llama3.1-8B-Ins	73.92	76.92	82.39
Qwen3-8B	79.92	80.06	84.29

Table 7. Test set Spearman R scores on influence regression task of three BERT-style models.

C.3. Difficulty Score Distribution

We list the difficulty score distributions of FineMed, Huatuo, Huatuo-DS, and UltraMedical in Figures 8, 9, 10, 11, and 12.

D. Details of Influence Score

D.1. Influence Score Computation

We directly computed influence scores for medium-scale datasets (FineMed, Huatuo, Huatuo-DS, m1, and MedReason) using Equation 4. However, to conserve computational resources, for the large-scale UltraMedical dataset, we instead trained an influence rater to predict these scores, following a similar strategy to our difficulty classifier. For this regression task, we evaluated BiomedBERT, ClinicalBERT, and ModernBERT. The models were trained on a set of 45,000 instances and tested on 5,000 instances, all sampled from the medium-scale datasets. As shown in Table 7, ModernBERT achieved the strongest performance and was selected as the influence rater.

D.2. Influence Score Distribution

We list the influence score distributions of FineMed, Huatuo, Huatuo-DS, m1, and MedReason in Figures 13, 14, 15, 16, and 17.

E. Details of Clinical Value Assessment

To ground our evaluation in clinical practice, we consulted three experienced clinicians to review the reasoning processes generated by our models. This expert review identified three components as crucial for establishing clinical value: *Differential Diagnosis (DDx)*, *Safety Check*, and *Evidence Citation*. These components subsequently formed the basis for our automated evaluation prompt, which is provided below.

Clinical Value

You are an experienced medical doctor and your task is to systematically evaluate and score the clinical reasoning process. The evaluation is structured around three core clinical cognitive behaviors: Differential Diagnosis (DDx), Safety Check, and Evidence Citation.

Question:

A 23-year-old woman presents to the emergency department complaining of a worsening headache. The headache started 1 month ago. It is constant and all over but gets worse when she is lying down or in the setting of bright lights. A review of systems is significant for a low-grade fever, night sweats, cough, malaise, poor appetite, and unintentional weight loss of 12 pounds in the last 2 months. The patient is sexually active with multiple male partners and reports inconsistent condom use. She has a history of intravenous drug use and has not been to a doctor in the last 2 years. Her temperature is 100.4 °F (38.0 °C), blood pressure is 110/78 mmHg, pulse is 88/min, and respirations are 14/min with an oxygen saturation of 98 % O₂ on room air. On physical exam, pain is elicited upon passive flexion of the patient's neck. A CT scan shows ventricular enlargement. A CD4⁺ count is 57 cells/μL blood. A lumbar puncture is performed with the following findings: **Opening pressure:** 210 mmH₂O; **Glucose:** 32 mg/dL; **Protein:** 204 mg/dL; **India ink stain:** Positive; **Leukocyte count and differential:** Leukocyte count: 200/mm³, Lymphocytes: 100 %. Which of the following treatments most directly addresses the underlying cause of this patient's symptoms?

- A. Acyclovir
- B. Amphotericin B and 5-flucytosine
- C. Dexamethasone
- D. Fluconazole
- E. Vancomycin and ceftriaxone

Answer: Okay, let's try to figure out this case. So, the patient is a 23-year-old woman with a worsening headache that's been going on for a month. The headache is constant, worse when lying down or in bright lights. She has a low-grade fever, night sweats, cough, malaise, poor appetite, and weight loss. She's sexually active with multiple partners and inconsistent condom use, and has a history of IV drug use. She hasn't seen a doctor in two years. Her vitals are temperature 100.4°F, BP 110/78, pulse 88, respirations 14, oxygen saturation 98%. On exam, neck pain with passive flexion. CT shows ventricular enlargement. *Evidence Citation:* CD4 count is 57 cells/μL. Lumbar puncture results: opening pressure 210 mmH₂O, glucose 32 mg/dL, protein 204 mg/dL. India ink stain is positive. CSF leukocytes 200/mm³, all lymphocytes, RBCs 2/mm³. First, the key symptoms and findings. The headache, neck stiffness (Brudzinski sign?), fever, weight loss, and the *Evidence Citation:* CD4 count is very low (57), which is indicative of advanced HIV. The *Evidence Citation:* India ink stain positive is a clue. India ink is used to detect Cryptococcus neoformans, which is a fungal infection. *Differential Diagnosis (DDx):* But wait, in someone with a low CD4 count, they're immunocompromised, so they could have opportunistic infections. But Cryptococcus is a common cause of meningitis in immunocompromised patients, especially those with HIV. But the *Evidence Citation:* CSF findings: elevated protein, low glucose, and India ink positive. So that's consistent with cryptococcal meningitis. The CSF leukocytes are 200, all lymphocytes, which is typical for cryptococcal meningitis. The red blood cells in CSF might be due to traumatic lumbar puncture, but the number is low here. The opening pressure is high, which is common in meningitis. *Safety Check:* But wait, let me check the options again. So the CSF findings are consistent with cryptococcal meningitis. Therefore, the correct treatment would be B. Amphotericin B and 5-flucytosine. However, in some cases, fluconazole might be used as initial treatment, but in more severe cases, the combination is preferred. Also, the patient's CD4 count is very low, so she might need antiretroviral therapy as well, but the question is about the treatment for the underlying cause of the symptoms, which is the cryptococcal infection. So B is the correct answer. Wait, but let me think again. The *Evidence Citation:* India ink stain is positive for Cryptococcus. So yes, the treatment is amphotericin B and 5-flucytosine. So the answer is B.

Figure 7. Case study of Qwen3-8B trained on DIQ-1% FineMed answering a question in MedBullets-option5.

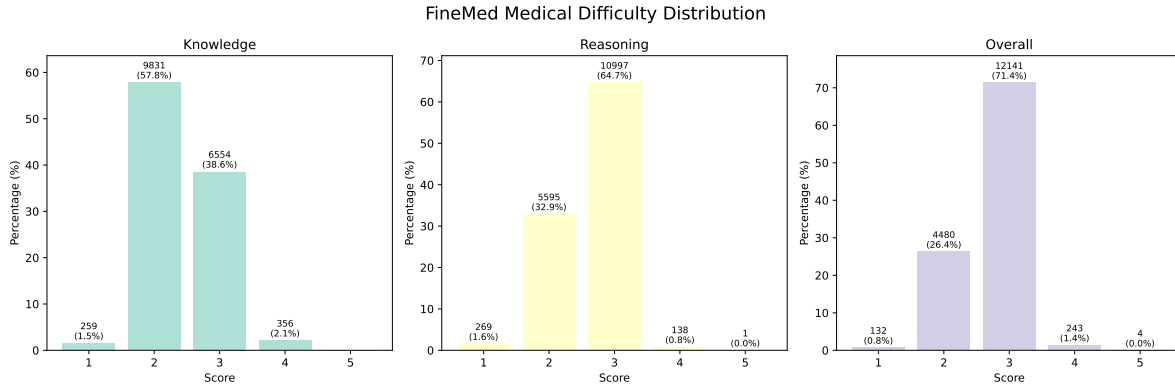


Figure 8. Difficulty score distribution of FineMed.

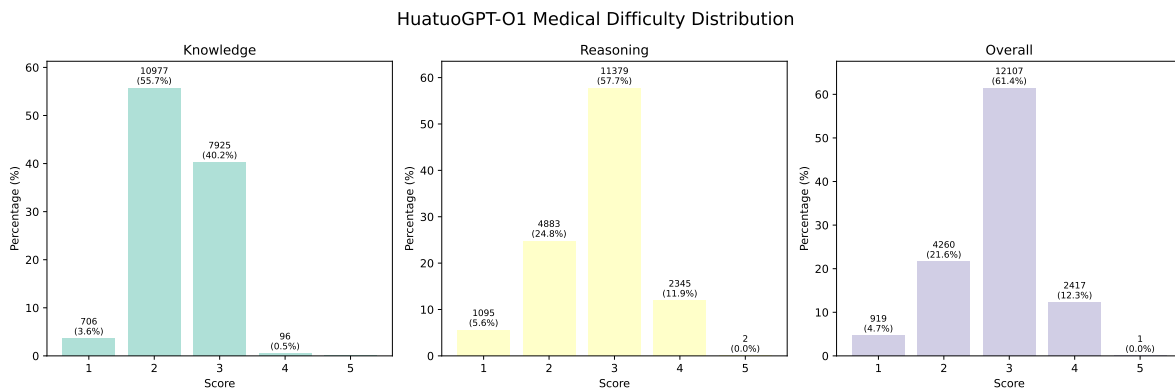


Figure 9. Difficulty score distribution of Huatuo.

Instructions:

- Read the medical Question and the Full Response: Carefully review the clinical scenario presented and the entire reasoning process.
- Evaluate Each Category Separately: For each of the three categories below, assess the performance against the described criteria.
- Assign a Score from 1 to 5: Use the detailed rubric to assign a score from 1 (Very Poor) to 5 (Excellent) for each category. Half-points (e.g., 3.5) are not permitted.
- Provide a Rationale: For each score, you must provide a brief, specific rationale explaining your decision. Justify the score by citing specific examples or omissions from the response.
- Use the Provided Output Template: Format your final evaluation using the template at the end of this document.

Scoring Rubric:

1. Differential Diagnosis (DDx)

This category assesses the ability to generate a list of potential diagnoses and systematically narrow it down using logical reasoning.

Level 5 (Excellent): The answer generates a comprehensive and relevant list of differential diagnoses, including both common and less common but critical possibilities. It systematically compares and contrasts the options, explaining why certain diagnoses are more or less likely. The process of elimination is clear, logical, and clinically astute, demonstrating a sophisticated understanding of disease presentation.

Level 4 (Good): The answer provides a relevant list of differential diagnoses and uses a logical process to narrow them down. The reasoning is clear and correct, though it may not explore the full spectrum of possibilities or the nuances between diagnoses as deeply as a Level 5 response.

Level 3 (Acceptable): The answer presents a limited but reasonable list of the most common differential diagnoses. It makes a plausible choice but the

HuatuogPT-O1-DS Medical Difficulty Distribution

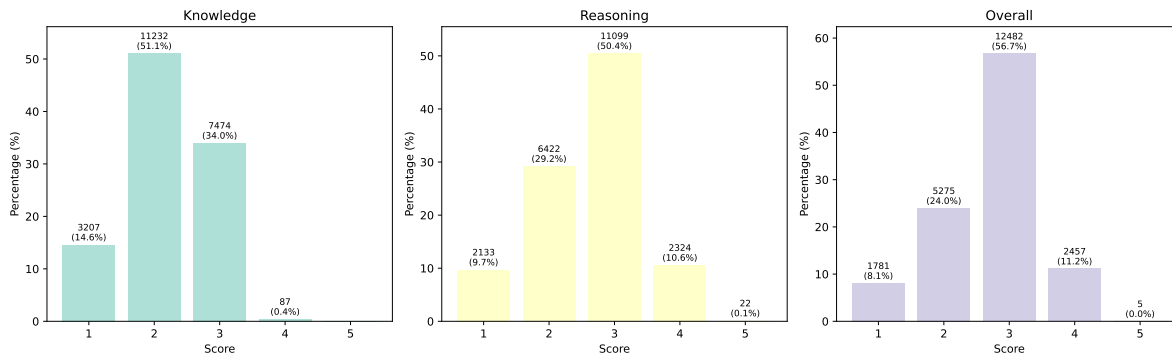


Figure 10. Difficulty score distribution of Huatuo-DS.

M1 Medical Difficulty Distribution

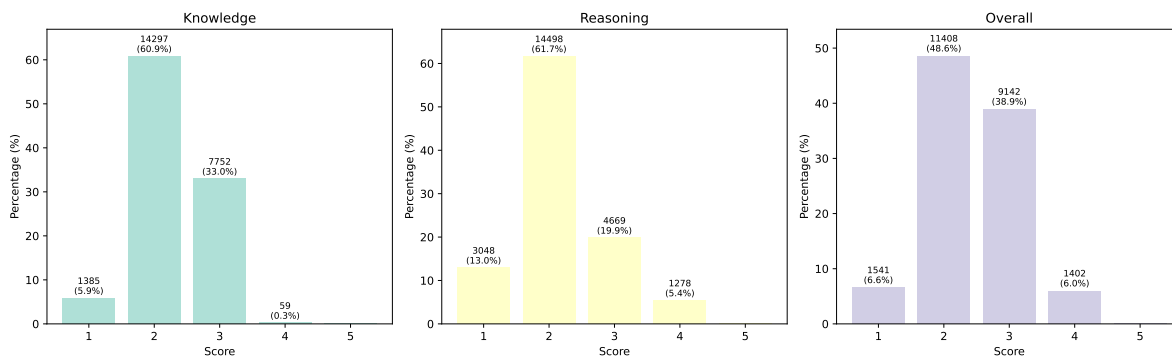


Figure 11. Difficulty score distribution of m1.

reasoning for excluding other options is superficial, weak, or absent. The process is functional but lacks depth.

Level 2 (Poor): The answer mentions one or two possible diagnoses but fails to create a structured list or engage in a meaningful comparison. It may jump to a conclusion prematurely or miss several obvious and important alternative diagnoses.

Level 1 (Very Poor): The answer fails to perform a differential diagnosis. It either provides a single answer with no consideration of alternatives or generates a list that is irrelevant, illogical, or factually incorrect.

2. Safety Check

This category assesses the ability to identify, prioritize, and mitigate potential risks to the patient. It reflects clinical responsibility and risk management.

Level 5 (Excellent): The answer demonstrates exceptional foresight. It not only identifies critical "red flag" conditions but also masterfully weighs complex, competing risks (e.g., balancing the risks

of a treatment against the risks of a disease). It correctly prioritizes the most immediate or severe threat and explains its risk-benefit analysis with clinical wisdom.

Level 4 (Good): The answer actively identifies and addresses significant safety concerns. It may discuss contraindications or weigh the pros and cons of different options from a safety perspective. The reasoning is proactive and demonstrates a strong awareness of patient safety.

Level 3 (Acceptable): The answer identifies and avoids obvious, direct risks or contraindications. It follows standard safety protocols (e.g., recommends confirming a diagnosis before treatment) but does not proactively analyze more complex or subtle risks. The behavior is reactive rather than proactive.

Level 2 (Poor): The answer misses a significant safety concern or mentions a risk but fails to act on it or incorporate it into the final decision. The safety awareness is present but insufficient for safe clinical practice.

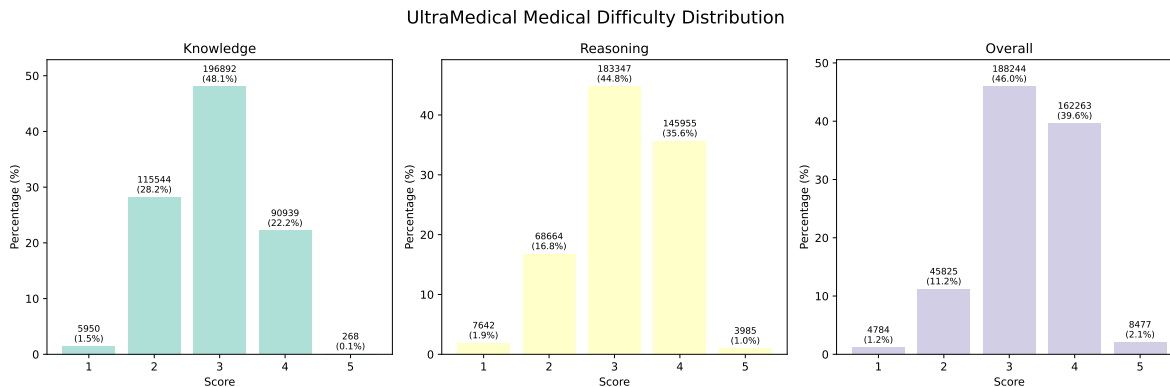


Figure 12. Difficulty score distribution of UltraMedical.

Level 1 (Very Poor): The answer makes a recommendation that is dangerous, contraindicated, or completely ignores a critical, life-threatening risk. The response poses a direct threat to patient safety.

3. Evidence Citation

This category assesses the ability to ground its reasoning in specific, relevant evidence, both from the patient’s data and from established medical knowledge.

Level 5 (Excellent): The answer seamlessly integrates multiple pieces of evidence (e.g., symptoms, lab values, patient history, and pharmacological data) into a cohesive and compelling argument. It not only cites evidence but also explains the significance and weight of key findings, demonstrating how specific evidence shifts the diagnostic probabilities. The reasoning is transparent and deeply rooted in evidence-based principles.

Level 4 (Good): The answer consistently and accurately cites relevant evidence to support its main arguments and conclusions. It clearly links specific findings (“Because of X...”) to its reasoning (“...we can conclude Y.”). It effectively uses a combination of patient-specific data and general medical facts.

Level 3 (Acceptable): The answer cites the most obvious pieces of evidence to support its final conclusion but may ignore other relevant data. The link between evidence and conclusion is present but may be simplistic. The reasoning is supported, but not robustly.

Level 2 (Poor): The answer mentions pieces of evidence from the prompt but fails to logically connect them to its reasoning or conclusion. The citation feels like a simple restatement of facts rather than an integrated part of an argument.

Level 1 (Very Poor): The answer makes claims without any supporting evidence, misinterprets the provided evidence, or uses irrelevant information to justify its conclusions.

Evaluation Output Template:

Please use the following format for your response.

- DDx Score: [1-5]
- Safety Check Score: [1-5]
- Evidence Citation Score: [1-5]
- Rationale for DDx: [Provide a brief, specific justification for the score, citing examples from the response.]
- Rationale for Safety Check: [Provide a brief, specific justification for the score, citing examples from the response.]
- Rationale for Evidence Citing: [Provide a brief, specific justification for the score, citing examples from the response.]

Here are the Question and Answer:

[QUESTION]

F. Details of Efficiency Analysis

We use Eq. 7 to approximate FLOPs for training on transformer-style models.

$$F_{\text{train}} = 6 \times L \times H^2 \times T \times |D_{\text{train}}| \times E \quad (7)$$

where L denotes the number of model layers, H denotes the hidden size, T denotes number of tokens per sample, $|D_{\text{train}}|$ denotes the number of training samples, and E denotes the number of training epochs. Similarly, the inference FLOPs can be approximated as:

$$F_{\text{infer}} = 2 \times L \times H^2 \times T \times |D_{\text{infer}}| \quad (8)$$

where $|D_{\text{infer}}|$ denotes the number of samples to infer on. For LoRA fine-tuning, the formula can be adapted to ac-

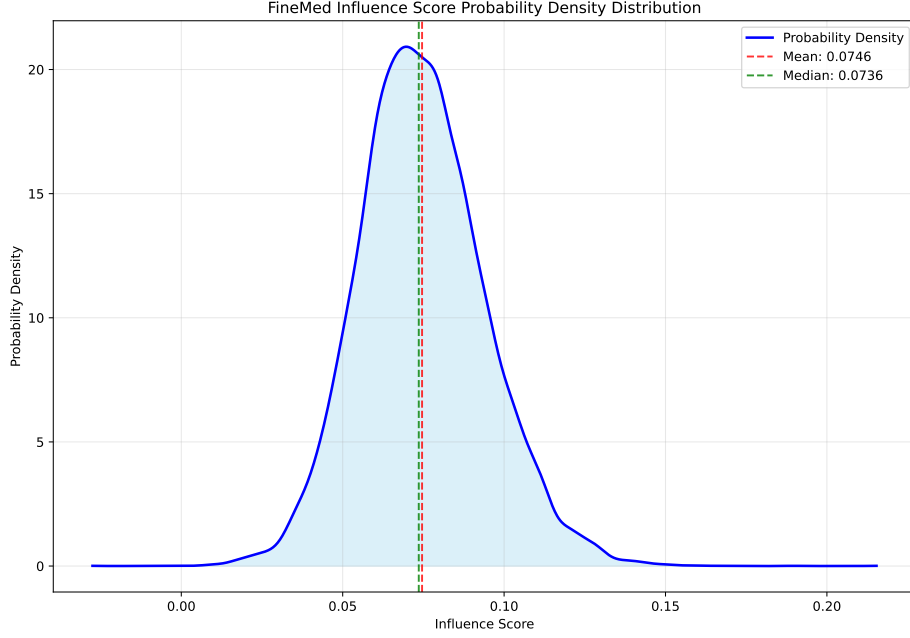


Figure 13. Influence score distribution of FineMed.

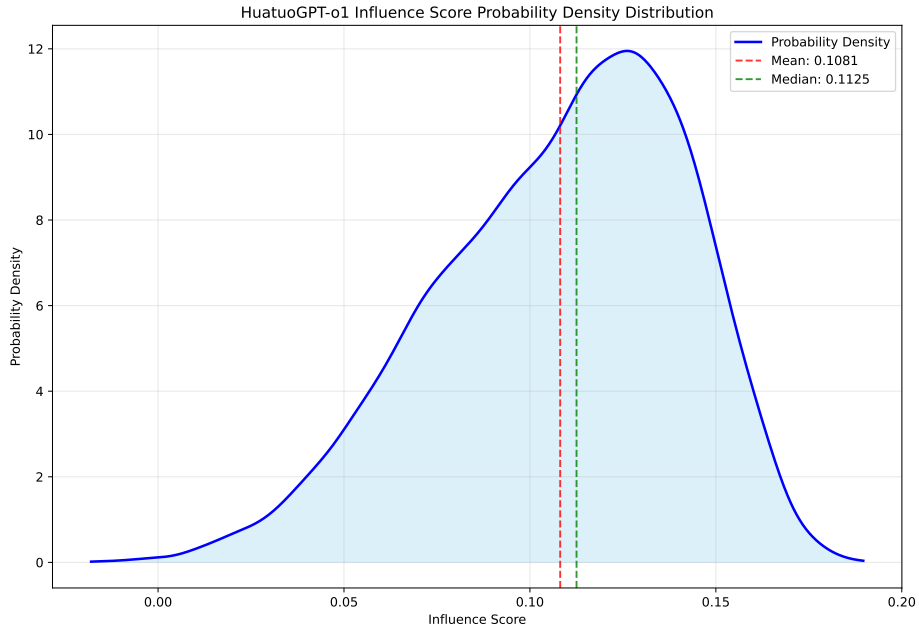


Figure 14. Influence score distribution of Huatuo.

count for the reduced number of trainable parameters. The core is replacing the quadratic dependency on the hidden size (H^2) with a term proportional to the LoRA rank (r) of the decomposition:

$$F_{\text{LoRA}} = 12 \times k \times L \times H \times r \times T \times |D_{\text{train}}| \times E \quad (9)$$

where r denotes the rank of the two LoRA matrices, and k denotes the number of matrices adapted with LoRA per layer ($k = 3$ in our experiment since LoRA is applied to query, key and value matrices in the self-attention blocks).

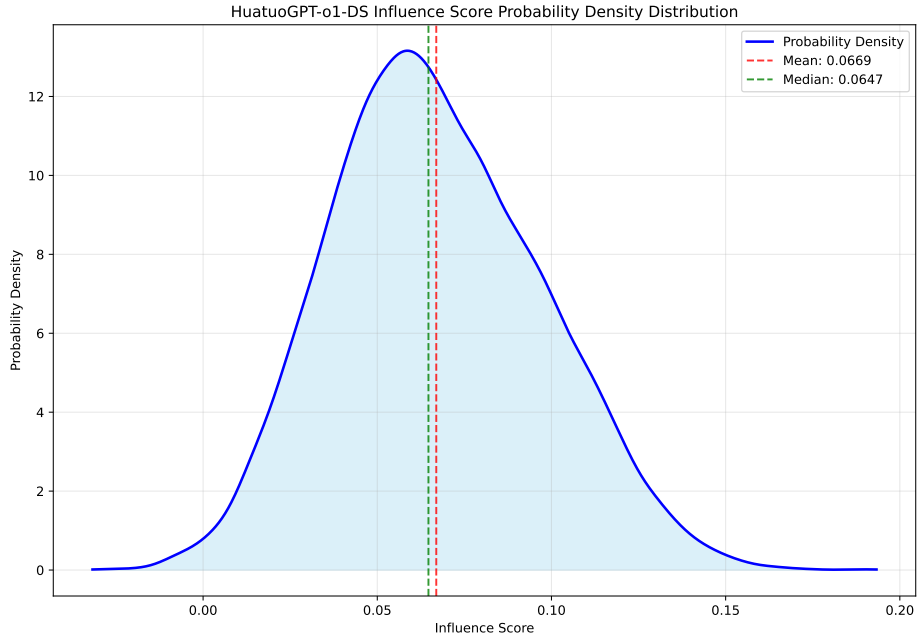


Figure 15. Influence score distribution of Huatuo-DS.

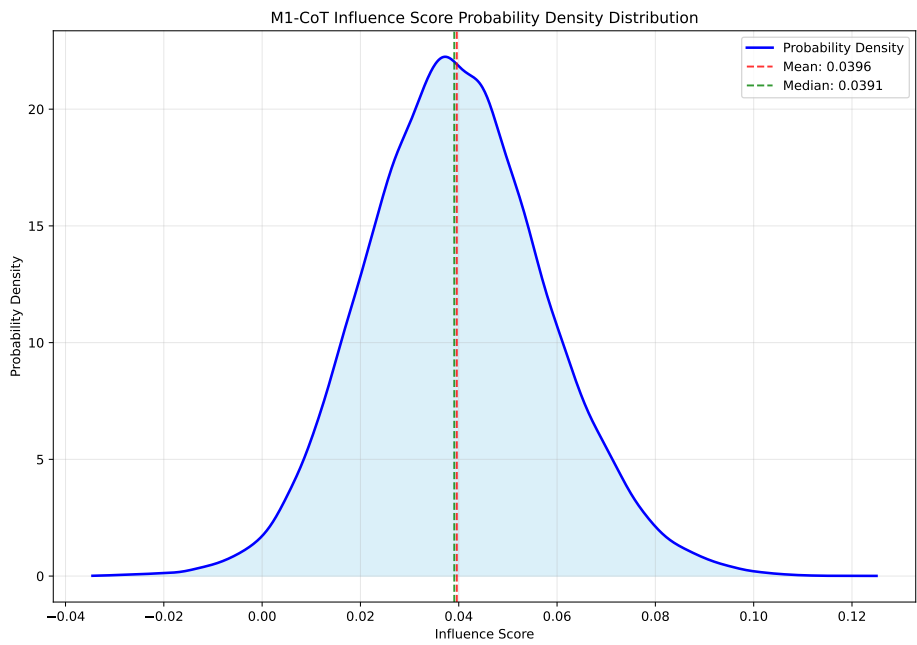


Figure 16. Influence score distribution of m1.

G. Full Experimental Results

9.

G.1. Main Results

Full downstream task results of general non-reasoning and reasoning models, and Llama-3.1-8B-Instruct trained on different selection setting data are provided in Tables 8 and

G.2. QA Experiment

Full downstream task results of QA experiment are provided in Table 10.

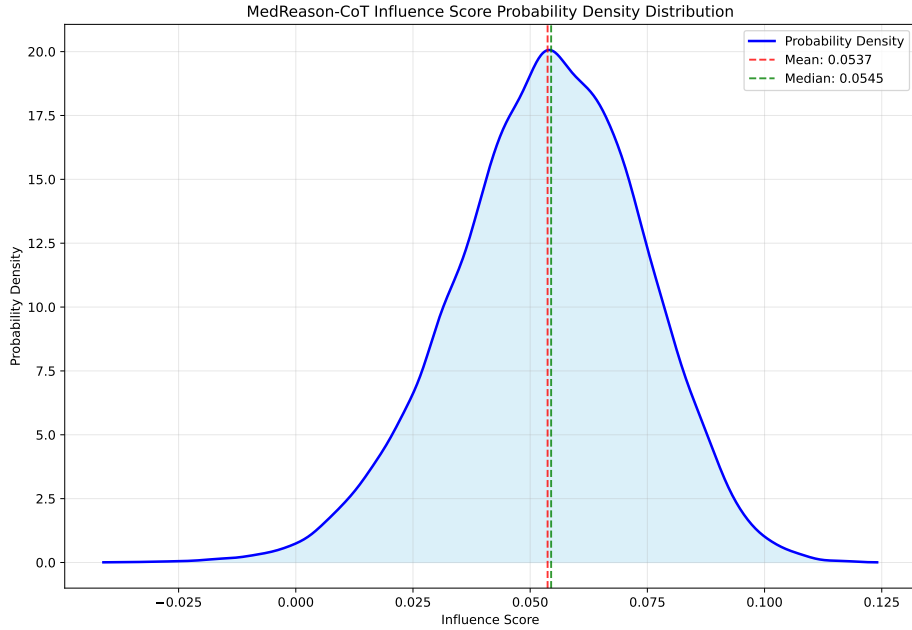


Figure 17. Influence score distribution of MedReason.

Model	MedQ	MedM	MMLU	Avg _S	HLE	MeB4	MeB5	MedX	MedG	MetM	Avg _C	Avg _A
GPT-4.1	84.29	73.34	82.46	80.03	7.77	71.75	70.13	42.00	64.44	70.79	54.48	63.00
DeepSeek-V3-0324	73.76	55.10	62.90	63.92	6.80	73.38	66.56	38.04	59.77	65.84	51.73	55.79
Gemini-2.5-flash	90.73	77.34	90.36	86.14	11.65	82.14	76.62	36.82	61.55	77.13	57.65	67.15
DeepSeek-R1-0528	92.85	76.55	91.28	86.89	13.59	83.44	54.22	38.61	59.88	72.91	53.78	64.81
QwQ-32B	75.10	63.45	78.97	72.51	12.62	67.86	59.09	22.65	48.44	63.80	45.74	54.66
o4-mini-medium	64.73	61.44	81.08	69.08	13.59	70.78	71.10	40.78	60.46	76.11	55.47	60.01
Gemini-2.5-pro	78.00	79.75	85.67	81.14	15.53	84.42	78.57	42.37	62.11	73.05	59.34	66.61

Table 8. Full downstream task results of general reasoning and non-reasoning models.

G.3. Ablation Study Experiment

Full downstream task results of DIQ ablation study are provided in Table 11.

G.4. Generalization Experiment

Full downstream task results of cross-scale and cross-model experiment, and preference learning experiment are provided in Tables 12 and 13.

Model	MedQ	MedM	MMLU	Avg _S	HLE	MeB4	MeB5	MedX	MedG	MetM	Avg _C	Avg _A
Llama3.1-8B-Inst	53.26	53.15	61.57	55.99	11.65	37.99	36.04	15.63	42.26	37.22	30.13	38.75
Huatuo	58.68	47.79	57.85	54.77	24.27	44.16	40.91	20.33	43.28	53.68	37.77	43.44
50% Random	57.74	48.39	57.94	54.69	23.30	46.75	40.26	18.49	41.76	35.54	34.35	41.13
50% DIQ	57.50	49.34	59.32	55.39	22.33	41.88	39.61	23.14	41.50	43.48	35.32	42.01
10% Random	52.87	47.33	56.75	52.32	15.53	39.94	30.84	17.51	40.71	35.32	29.98	37.42
10% DIQ	58.13	53.57	62.63	58.11	25.24	44.48	40.40	17.59	43.38	50.91	37.00	44.04
1% Random	54.75	43.99	55.19	51.31	15.86	42.50	37.71	13.63	44.75	46.39	33.47	39.42
1% DIQ	56.64	50.16	62.81	56.54	13.59	47.40	47.75	14.45	45.86	46.39	35.91	42.78
Huatuo-DS	67.24	58.26	74.38	66.63	11.65	57.47	51.62	15.71	43.01	53.68	38.86	48.11
50% Random	64.34	55.77	71.72	63.94	7.77	53.90	51.62	15.59	42.71	53.82	37.57	46.36
50% DIQ	66.46	56.49	72.18	65.04	12.62	54.22	50.65	16.45	42.78	53.17	38.32	47.22
10% Random	65.12	57.38	71.26	64.59	12.62	57.14	48.05	14.41	41.65	52.08	37.66	46.63
10% DIQ	67.87	57.57	73.00	66.15	14.56	59.35	50.97	15.55	44.89	54.84	40.03	48.73
1% Random	56.95	49.63	62.53	56.37	7.77	45.86	41.88	13.18	40.29	41.66	31.77	39.97
1% DIQ	63.16	53.93	63.36	60.15	16.50	46.10	44.48	25.47	39.27	32.19	37.96	45.36
FineMed	40.22	51.26	51.61	47.70	16.50	46.10	44.48	25.47	39.27	32.19	34.00	38.57
50% Random	40.38	33.83	37.47	37.23	17.48	42.86	43.83	21.43	40.12	38.53	34.04	35.10
50% DIQ	42.66	35.24	39.49	39.13	18.45	50.65	43.83	19.47	41.87	37.14	35.24	36.53
10% Random	51.14	39.04	45.27	45.15	16.50	45.13	42.50	16.49	42.89	40.93	34.07	37.77
10% DIQ	51.61	40.40	45.91	45.97	17.48	48.05	43.83	18.57	44.87	43.55	36.06	39.36
1% Random	51.61	48.98	58.68	53.09	11.65	45.45	42.86	13.59	40.29	35.76	31.60	38.76
1% DIQ	53.50	54.15	66.76	58.14	12.62	45.45	42.21	13.80	44.28	40.35	33.12	41.46
m1	75.88	64.33	82.83	74.35	16.50	66.56	60.06	17.35	43.68	58.99	43.86	54.02
50% Random	74.31	62.68	82.46	73.15	14.56	64.29	58.44	19.39	41.64	61.62	43.32	53.27
50% DIQ	74.86	63.47	82.19	73.51	17.48	64.61	56.82	17.76	43.35	61.54	43.59	53.56
10% Random	75.26	60.53	79.98	71.81	12.62	60.39	57.79	17.59	46.97	60.52	42.65	52.41
10% DIQ	74.39	61.25	79.80	71.92	16.50	62.66	58.44	17.59	46.61	62.49	44.05	53.30
1% Random	71.09	58.52	77.78	69.13	4.85	55.84	53.25	16.41	46.50	58.49	39.22	49.19
1% DIQ	72.03	60.51	76.95	69.83	10.68	61.36	54.87	16.82	46.31	59.72	41.63	51.03
MedReason	61.51	42.34	73.28	59.04	9.71	34.87	36.17	15.96	43.38	37.22	29.55	39.38
50% Random	60.41	50.39	66.48	59.09	11.65	36.69	32.14	18.29	48.92	34.01	30.28	39.89
50% DIQ	63.24	53.07	71.07	62.46	16.50	36.69	28.90	18.29	48.48	35.40	30.71	41.29
10% Random	58.99	51.61	64.46	58.35	11.65	33.44	26.30	15.84	42.27	44.57	29.01	38.79
10% DIQ	59.62	49.18	59.41	56.07	22.33	44.81	44.48	17.47	39.74	48.94	36.30	42.89
1% Random	45.40	35.38	41.41	40.73	18.45	40.58	37.99	15.35	42.74	46.39	33.58	35.97
1% DIQ	46.11	44.59	51.61	47.44	21.36	41.23	38.31	17.80	48.48	46.39	35.60	39.54
UltraMedical	67.24	51.95	69.70	62.96	23.30	51.95	50.32	15.22	43.63	45.45	38.31	46.53
50% Random	67.09	49.59	67.31	61.33	14.56	52.57	46.43	13.18	44.60	42.83	35.70	44.24
50% DIQ	67.09	53.50	69.42	63.34	18.45	53.25	51.30	13.80	44.95	43.92	37.61	46.19
10% Random	66.22	55.73	71.35	64.43	10.68	56.49	46.78	13.80	48.02	43.70	36.58	45.86
10% DIQ	67.79	54.41	72.82	65.01	12.62	56.57	47.40	15.18	48.24	44.28	37.38	46.59
1% Random	65.28	54.77	71.72	63.92	12.62	53.90	45.78	13.88	47.40	43.26	36.14	45.40
1% DIQ	66.69	58.52	71.63	65.61	11.65	57.14	46.75	13.59	47.21	44.65	36.83	46.43

Table 9. Full downstream task results of trained Llama-3.1-8B-Instruct using full dataset, random subset, and our DIQ selected data at 1%, 10%, and 50% data keeping ratios.

Setting	MedQ	MedM	MMLU	Avg _S	HLE	MeB4	MeB5	MedX	MedG	MetM	Avg _C	Avg _A
Llama3.1-8B-Instruct												
m1-23k-QA												
50% Random	67.64	58.40	75.85	67.30	8.74	57.47	52.27	14.90	46.65	53.90	38.99	48.42
50% DIQ	66.93	58.57	75.85	67.12	19.42	54.22	52.92	13.80	48.99	55.35	40.78	49.56
10% Random	65.91	58.36	75.67	66.65	10.68	55.84	48.38	13.63	46.88	54.19	38.27	47.73
10% DIQ	66.06	57.73	75.21	66.33	20.39	55.52	50.32	14.08	46.91	54.84	40.34	49.01
1% Random	60.09	51.11	64.83	58.68	12.62	47.08	43.51	18.16	49.15	46.25	36.13	43.64
1% DIQ	62.69	53.41	64.92	60.34	16.50	52.27	46.43	16.73	48.53	45.88	37.72	45.26
MedReason-QA												
50% Random	58.99	51.49	72.73	61.07	20.39	36.36	33.12	17.14	48.84	38.75	32.43	41.98
50% DIQ	59.62	54.58	75.76	63.32	20.39	45.45	38.96	23.35	47.52	44.94	36.77	45.62
10% Random	61.19	50.75	67.13	59.69	17.48	37.66	34.74	17.96	48.81	40.93	32.93	41.85
10% DIQ	60.64	54.27	72.54	62.48	29.13	53.90	49.35	18.65	39.76	48.22	39.84	47.38
1% Random	44.62	36.53	46.19	42.45	24.27	43.51	39.61	14.57	46.17	38.24	34.40	37.08
1% DIQ	45.40	42.96	53.08	47.15	26.21	43.18	39.94	17.18	47.04	46.54	36.68	40.17
Qwen3-8B												
m1-23k-QA												
50% Random	64.41	55.87	78.70	66.33	13.59	52.60	44.81	15.59	45.28	51.64	37.25	46.94
50% DIQ	65.75	55.82	77.87	66.48	18.45	57.14	46.75	16.86	47.92	50.47	39.60	48.56
10% Random	64.02	57.85	78.79	66.89	13.59	52.92	43.51	15.39	48.06	52.44	37.65	47.40
10% DIQ	65.75	57.85	79.71	67.77	17.48	49.03	46.75	15.22	49.06	53.02	38.43	48.21
1% Random	77.14	63.54	84.11	74.93	16.50	64.94	57.47	15.22	54.45	61.76	45.06	55.01
1% DIQ	77.14	64.26	85.40	75.60	20.39	65.91	61.69	23.35	54.96	62.56	48.14	57.30
MedReason-QA												
50% Random	56.32	51.90	75.57	61.26	19.42	41.56	33.12	16.29	50.16	37.58	33.02	42.44
50% DIQ	58.29	53.62	74.84	62.25	25.24	40.58	42.86	16.69	48.44	40.35	35.69	44.55
10% Random	52.40	50.94	66.76	56.70	19.42	40.91	35.06	14.73	47.57	28.91	31.10	39.63
10% DIQ	56.64	52.98	69.88	59.83	19.42	49.03	45.45	17.96	48.47	51.27	38.60	45.68
1% Random	35.19	41.55	53.81	43.52	7.77	39.94	30.52	13.02	29.87	35.83	26.16	31.94
1% DIQ	52.95	46.98	64.46	54.80	14.56	44.16	41.56	14.12	43.27	45.96	33.94	40.89

Table 10. Full downstream task results of Llama3.1-8B-Instruct and Qwen3-8B under training on 1%, 10%, and 50% randomly selected and DIQ-selected QA datasets.

Setting	MedQ	MedM	MMLU	Avg _S	HLE	MeB4	MeB5	MedX	MedG	MetM	Avg _C	Avg _A
Llama3.1-8B-Instruct												
50% Influence	56.95	50.11	56.75	54.60	22.33	45.45	36.36	21.96	41.72	41.30	34.85	41.44
50% Overall	58.21	48.86	56.75	54.61	22.33	44.16	37.66	17.76	44.33	35.69	33.66	40.64
50% Knowledge	57.97	48.12	60.51	55.53	26.21	43.51	36.69	18.33	41.02	36.78	33.76	41.02
50% Reasoning	57.11	47.43	56.20	53.58	21.36	51.30	38.31	18.29	43.57	34.89	34.62	40.94
50% DIQ	57.50	49.34	59.32	55.39	22.33	41.88	39.61	23.14	41.50	43.48	35.32	42.01
10% Influence	58.37	51.78	61.98	57.38	25.24	40.26	38.31	20.20	42.35	45.01	35.23	42.61
10% Overall	54.36	51.30	59.69	55.12	23.30	46.10	40.26	20.24	42.25	50.91	37.18	43.16
10% Knowledge	53.42	49.37	56.38	53.06	23.30	43.18	35.06	18.86	40.79	39.69	33.48	40.01
10% Reasoning	55.15	50.63	62.35	56.04	25.24	42.53	37.66	17.76	42.17	43.48	34.81	41.89
10% DIQ	58.13	53.57	62.63	58.11	25.24	44.48	40.40	17.59	43.38	50.91	37.00	44.04
1% Influence	55.85	46.71	62.73	55.10	12.62	51.82	47.73	13.18	45.17	48.22	36.46	42.67
1% Overall	53.10	42.46	57.94	51.17	16.50	47.73	47.73	13.14	45.75	47.92	36.46	41.36
1% Knowledge	54.20	44.97	57.30	52.16	18.45	51.62	42.21	12.69	44.95	47.92	36.31	41.59
1% Reasoning	55.70	44.63	58.68	53.00	7.77	50.65	42.86	13.51	44.79	47.71	34.55	40.70
1% DIQ	56.64	50.16	62.81	56.54	13.59	47.40	47.75	14.45	45.86	46.39	35.91	42.78
Qwen3-8B												
50% Influence	60.33	53.86	74.38	62.86	20.39	45.13	44.48	32.78	45.28	44.72	38.80	46.82
50% Overall	59.39	53.05	74.10	62.18	16.50	46.10	41.23	31.96	45.31	41.08	37.03	45.41
50% Knowledge	59.78	54.46	73.00	62.41	13.59	41.88	39.61	29.76	44.78	40.50	35.02	44.15
50% Reasoning	61.82	53.74	73.09	62.88	21.36	46.75	39.61	30.61	44.02	40.79	37.19	45.75
50% DIQ	62.45	54.08	73.28	63.27	14.56	51.30	47.73	33.67	46.03	46.61	39.98	47.75
10% Influence	64.81	54.84	75.94	65.20	14.56	49.35	41.23	15.06	44.82	49.75	35.80	45.60
10% Overall	63.55	54.20	74.75	64.17	15.53	51.30	44.81	15.43	43.38	44.94	35.90	45.32
10% Knowledge	63.71	53.91	75.67	64.43	18.41	47.40	46.75	17.10	44.51	48.36	37.09	46.20
10% Reasoning	63.55	53.69	75.11	64.12	12.62	47.40	43.83	15.71	43.63	46.76	34.99	44.70
10% DIQ	65.51	56.30	76.95	66.25	13.59	51.62	48.70	18.29	47.04	50.62	38.31	47.62
1% Influence	77.85	63.81	82.28	74.65	13.59	63.64	57.14	18.69	54.81	60.96	44.81	54.75
1% Overall	77.53	64.04	81.63	74.40	14.56	64.61	55.19	18.78	54.74	61.76	44.94	54.76
1% Knowledge	78.00	63.52	82.92	74.81	11.65	61.04	57.14	18.24	54.74	61.62	44.07	54.32
1% Reasoning	76.28	63.69	83.10	74.36	10.68	64.94	57.47	17.71	55.13	61.18	44.52	54.46
1% DIQ	77.45	63.93	82.74	74.71	15.53	66.88	58.12	18.53	54.74	61.76	45.93	55.52

Table 11. Full downstream task results of Llama3.1-8B-Instruct and Qwen3-8B trained using datasets selected by different selection scores, under data keeping ratio of 1%, 10%, and 50%.

Model	MedQ	MedM	MMLU	Avg _S	HLE	MeB4	MeB5	MedX	MedG	MetM	Avg _C	Avg _A
Qwen3-8B												
50% Random	60.57	54.41	68.41	61.13	22.33	48.38	45.13	25.47	44.33	37.29	37.16	45.15
50% DIQ Llama Inf	63.32	53.36	77.23	64.64	15.53	50.00	44.81	26.65	44.84	43.77	37.60	46.61
50% DIQ Qwen Inf	62.45	54.08	73.28	63.27	14.56	51.30	47.73	33.67	46.03	46.61	39.98	47.75
10% Random	63.94	52.21	76.03	64.06	12.62	47.40	44.81	21.88	46.16	44.57	36.24	45.51
10% DIQ Llama Inf	65.51	56.30	76.95	66.25	13.59	51.62	48.70	18.29	47.04	50.62	38.31	47.62
10% DIQ Qwen Inf	67.01	55.25	77.96	66.74	12.62	55.19	47.40	17.71	46.79	52.29	38.67	48.02
1% Random	76.04	61.08	82.19	73.10	9.71	65.58	54.55	16.12	54.45	57.83	43.04	53.06
1% DIQ Llama Inf	76.28	63.93	82.74	74.32	13.59	66.88	57.47	17.71	54.74	60.96	45.23	54.92
1% DIQ Qwen Inf	76.67	64.45	83.38	74.83	15.53	64.29	57.14	18.12	54.96	61.18	45.20	55.08
Qwen3-14B												
50% Random	64.02	57.69	73.74	65.15	17.48	62.34	52.92	30.41	48.60	58.19	44.99	51.71
50% DIQ Llama Inf	65.12	58.55	72.73	65.47	23.30	59.42	56.49	32.20	49.58	60.16	46.86	53.06
50% DIQ Qwen Inf	67.32	58.45	74.20	66.66	19.42	62.99	58.12	29.31	49.70	59.94	46.58	53.27
10% Random	65.12	57.76	77.41	66.76	13.59	59.09	56.17	27.10	48.20	53.53	42.95	50.89
10% DIQ Llama Inf	66.38	58.16	78.42	67.65	10.68	59.09	53.57	20.61	47.26	56.23	41.36	50.04
10% DIQ Qwen Inf	71.64	59.50	81.54	70.89	17.48	63.64	56.49	20.12	48.37	56.23	43.72	52.78
1% Random	80.99	66.99	84.94	77.64	9.71	69.06	61.69	20.12	54.02	63.07	46.69	56.73
1% DIQ Llama Inf	81.38	67.32	85.95	78.22	6.80	69.06	61.69	20.20	54.90	63.15	45.97	56.72
1% DIQ Qwen Inf	82.09	67.73	85.31	78.38	10.68	73.38	64.29	20.61	55.25	64.02	47.42	58.15
Qwen3-32B												
50% Random	66.93	61.61	77.50	68.68	16.50	60.06	50.32	23.67	49.30	56.15	42.67	51.34
50% DIQ Llama Inf	69.21	61.25	74.38	68.28	22.33	55.52	53.90	23.71	50.41	60.16	44.34	52.32
50% DIQ Qwen Inf	69.68	61.01	76.40	69.03	29.13	58.44	55.84	26.29	50.17	58.85	46.45	53.98
10% Random	67.64	58.76	76.58	67.66	18.45	65.58	62.34	26.98	49.66	62.05	47.51	54.23
10% DIQ Llama Inf	70.46	59.48	76.95	68.96	18.45	66.56	62.01	28.12	50.54	60.23	47.65	54.76
10% DIQ Qwen Inf	71.80	59.41	75.67	68.96	25.24	64.61	66.88	32.29	50.72	63.15	50.48	56.64
1% Random	79.73	60.22	74.84	71.60	10.68	70.13	66.56	28.41	58.20	63.36	49.56	56.90
1% DIQ Llama Inf	78.38	59.05	78.42	71.95	13.59	68.83	64.29	24.08	57.43	64.31	48.76	56.49
1% DIQ Qwen Inf	78.24	61.25	81.54	73.68	16.50	66.88	64.61	26.57	57.33	63.22	49.19	57.35

Table 12. Full downstream task results of Qwen series models trained using DIQ under different influence score settings.

Model	MedQ	MedM	MMLU	Avg _S	HLE	MeB4	MeB5	MedX	MedG	MetM	Avg _C	Avg _A
FineMed	74.34	63.23	82.29	73.29	13.59	66.22	58.27	21.02	55.31	61.11	45.92	55.04
FineMed + DPO	75.51	65.30	83.67	74.83	14.64	66.56	58.44	18.69	55.31	61.55	45.87	55.52
50% Random + DPO	78.16	63.71	82.74	74.87	9.71	66.56	58.44	18.61	55.31	61.11	44.96	54.93
50% DIQ + DPO	77.85	64.57	81.91	74.78	10.68	68.51	58.44	19.59	54.86	61.91	45.67	55.37
10% Random + DPO	76.75	63.88	83.47	74.70	8.74	65.26	60.39	18.69	54.67	62.86	45.10	54.97
10% DIQ + DPO	76.51	64.74	82.28	74.51	14.64	65.91	62.42	17.64	54.77	62.35	46.29	55.70
1% Random + DPO	78.08	64.33	81.82	74.74	17.48	65.26	59.42	18.90	54.98	62.13	46.36	55.82
1% DIQ + DPO	76.98	64.57	84.02	75.19	16.50	65.58	64.29	18.90	54.96	62.86	47.18	56.52

Table 13. Full downstream task results of Qwen3-8B fine-tuned using DIQ-selected or random data and DPO.