

000 SINGLE ANSWER IS NOT ENOUGH: ON GENERATING 001 RANKED LISTS WITH MEDICAL REASONING MODELS 002

003 **Anonymous authors**

004 Paper under double-blind review

005 ABSTRACT

006 This paper presents a systematic study on enabling *medical* reasoning models
007 (MRMs)—which achieve SOTA performance on multiple-choice benchmarks—to
008 remain robust when producing alternative *answer formats*. Answer formats define
009 the structure of a final answer in a generated response, such as an option, free text,
010 or a ranked list. Although clinical decision-making typically involves weighing
011 multiple plausible possibilities, current MRMs are trained to produce only one
012 answer, and their robustness beyond that format is not well studied. We focus
013 on the *ranked-list* format as an alternative that better reflects clinical uncertainty.
014 To address this gap, we evaluate *prompting* and *fine-tuning* for enabling MRMs
015 to generate ranked lists across common medical benchmarks. While prompting
016 provides a lightweight solution, MRMs vary widely in their ability to follow such
017 instructions. We therefore explore supervised fine-tuning (SFT) and reinforcement
018 fine-tuning (RFT) as stronger adaptation methods. SFT trains models to imitate
019 ranked outputs, whereas RFT optimizes behavior through reward functions; we
020 introduce new rewards tailored to ranked-list generation and analyze their effects
021 through ablations. Our results show that although some SFT models handle cer-
022 tain formats well, RFT yields more consistent robustness across multiple answer
023 formats. A case study on a modified MedQA benchmark with multiple valid an-
024 swers further reveals that MRMs can recognize clinically sound alternatives even
025 when misaligned with a benchmark’s preferred ground truth. To the best of our
026 knowledge, this is the first systematic investigation of adapting MRMs to alter-
027 native answer formats such as ranked lists. We hope this study lays the foundation
028 for developing more flexible and clinically aligned MRMs.

What is the most appropriate initial diagnostic test for a pregnant woman presenting with severe, episodic back pain radiating to the abdomen and hematuria?

?

Short Text Answer

As an expert clinical assistant, I can provide information on standard medical practices.

For a pregnant woman presenting with severe, episodic back pain radiating to the abdomen and hematuria, the clinical suspicion is high for nephrolithiasis (kidney stones)

Why Not Other Tests (Initially)?

- Non-contrast CT Scan: ...
- MRI/MR Urography: ...
- KUB X-ray: ...

In addition to the ultrasound, a urinalysis and urine culture would be performed ... urinary tract infection.

The most appropriate initial diagnostic test for a pregnant woman with symptoms suggestive of a kidney stone is a renal and bladder ultrasound.

The most appropriate initial diagnostic test for a pregnant woman with symptoms suggestive of a kidney stone is a renal and bladder ultrasound.

\boxed{\text{Renal and bladder ultrasound}} ✓✓✓

Ranked List Answer

<think> Considering the symptoms of severe, episodic back pain radiating to the abdomen and hematuria in a pregnant woman, the most likely condition is a urinary tract infection (UTI) or a kidney stone. The initial diagnostic test for kidney stones in this context would be a urinary sediment analysis, followed by a renal ultrasound if necessary. Other options like urine culture, CT scan, or plain x-ray are less initial and more invasive.</think>

Final Answer

1. Urine dipstick test (urinalysis) ✓✓✓
2. Renal ultrasound ✓✓✓
3. Urine culture ✓✓
4. Intravenous pyelogram (IVP) ✓
5. Abdominal plain x-ray ✓

Answer Formats

	Choice
	Short Text
	Ranked List

Approaches

	Prompting
	Supervised Fine-Tuning
	Reinforcement Fine-Tuning

050 Figure 1: **Left** Comparison of model responses between a correct *short answer* and a *list answer*
051 containing possible correct answers, which depends on the clinical context. **Right** We systematically
052 study medical LLMs across three answer formats—*choice*, *short text*, and *list answers*—using three
053 approaches: *prompting*, *supervised fine-tuning*, and *reinforcement fine-tuning*

054 **1 INTRODUCTION**

055

056 Recent advances in reasoning models have shown remarkable progress in verifiable domains such
 057 as mathematics and programming (Comanici et al., 2025; Yang et al., 2025; Guo et al., 2025). The
 058 medical domain, however, presents a distinct challenge. Clinical decision-making rarely involves
 059 selecting a single best treatment; instead, it requires managing uncertainty by weighing multiple
 060 plausible options (Varkey, 2020). Current medical reasoning models (MRMs), trained primarily
 061 on multiple-choice question (MCQ) datasets, inherit a single-best-answer bias and are optimized to
 062 produce only one answer (Chen et al., 2024; Huang et al., 2025a; Liu et al., 2025a).

063

064 In practice, single-answer outputs can mislead clinicians, who may over-rely on automated, potentially
 065 incorrect recommendations. By contrast, presenting a ranked list of options better reflects
 066 diagnostic uncertainty and supports safer decision-making (Goddard et al., 2012; Meng et al., 2024;
 067 Tao et al., 2020). For example, Jabbour et al. (2023) show that clinicians are more likely to adopt an
 068 incorrect diagnosis of pneumonia—even when it contradicts their initial judgment—simply because it
 069 appears as the model’s proposed answer. This risk naturally extends to MRMs that produce only
 070 single-answer responses.

071 We argue that a *ranked list of plausible options* is a more informative and clinically aligned answer
 072 format. Ranked lists mirror clinical practice, where practitioners generate a differential diagnosis
 073 (DDx) (Lamba et al., 2021) before forming management plans, and they encourage collaboration
 074 between models and human experts by prompting reevaluation of the presented options. Figure 1
 075 illustrates the contrast between single-answer and ranked-list responses.

076 Despite these motivations, the **robustness**¹ of MRMs across **answer formats** remains understudied—
 077 specifically, *how well a model trained in one format can follow instructions to produce another*.
 078 For example, it is unclear how reliably an MCQ-trained MRM can generate a ranked list when
 079 prompted. Our study focuses on ranked lists with a single ground-truth answer, evaluating whether
 080 the correct answer appears at the top of the list. This setup reflects limitations of existing medical QA
 081 benchmarks, as discussed in Section 5, and differs from differential diagnosis tasks in which multiple
 082 plausible answers matter (Lim et al., 2025). Nevertheless, our goal extends beyond differential
 083 diagnosis: we seek to understand answer-format robustness more broadly across general medical
 084 QA and related settings.

085 This motivates the central question of our study: **How can MRMs be made to generate ranked**
 086 **lists from open-ended problems?** To answer this, we investigate two approaches: **prompting**,
 087 which guides existing models toward the desired format, and **fine-tuning**, which trains models to
 088 produce responses in that format.

089 Prompting offers a lightweight way to steer model behavior without additional training; however,
 090 its effectiveness for MRMs is less understood than in general medical LLMs (Sahoo et al., 2025;
 091 Schulhoff et al., 2025). Compounding this challenge, recent work suggests that reasoning models
 092 may exhibit degraded instruction-following ability (Li et al., 2025a; Fu et al., 2025; Jang et al.,
 093 2025), directly limiting their ability to adopt new answer formats. We therefore examine how well
 094 prompting alone can steer MCQ-trained MRMs to generalize to other formats.

095 Fine-tuning, in contrast, modifies model weights and provides a stronger mechanism for steering
 096 behavior. We study two established methods: supervised fine-tuning (SFT) and reinforcement fine-
 097 tuning (RFT). SFT trains models on annotated prompt–response pairs, which in reasoning models
 098 can be derived through data synthesis (Chen et al., 2024; Qin et al., 2024) or knowledge distillation
 099 (Huang et al., 2024; 2025b). RFT instead optimizes reward signals (Zhang et al., 2025a; Guo et al.,
 100 2025), allowing models to discover behaviors that better align with the target answer format. We
 101 evaluate how MRMs trained via SFT or RFT generalize across answer formats and conduct ablation
 102 studies to analyze how different RFT design choices affect performance and stability.

103 Our findings show that SFT-trained MRMs can inadvertently entangle answer format with knowledge
 104 recall, reducing performance when evaluated on formats not seen during training. SFT models
 105 trained specifically on ranked lists generalize poorly to other formats. In contrast, RFT models

106
 107 ¹In this paper, we use the terms *robust* and *generalizes* interchangeably to describe a model’s ability to
 108 perform well across different answer formats.

108 trained on the MCQ format exhibit strong robustness across answer formats, though they still underperform compared with RFT models trained directly on the ranked-list format.
 109
 110

111 Our contributions are as follows:

112 • We present the first systematic study of steering MRMs via prompting and fine-tuning for ranked-
 113 list generation. We introduce an evaluation framework covering three answer formats—MCQ, QA,
 114 and List—and a pipeline for converting MCQ benchmarks into open-ended formats. We find that
 115 many models may know the correct answers but may often fail to select the benchmark-preferred
 116 one.
 117 • We conduct a comprehensive study of SFT and RFT under specific answer formats for MRMs.
 118 We show that SFT trained on MCQs generalizes reasonably well across formats, whereas RFT
 119 models generalize both to unseen examples and across answer formats.
 120 • We develop and propose reward functions for RFT targeting ranked-list generation, each offering
 121 different trade-offs in model behavior.
 122 • We perform extensive ablation studies of RFT for MRMs, demonstrating that sequencing multiple
 123 RFT stages as a curriculum over answer formats (MCQ → List) improves stability and reduces
 124 collapse compared with training directly on the list format.

125 2 EXPERIMENTAL SETUP

126 2.1 ANSWER FORMATS

127 We define an *answer format* as the structure in which the model must present the final answer in
 128 its generated response. In this study, we focus on three formats: **MCQ**, **QA**, and **List**. The **MCQ**
 129 format evaluates models in the conventional multiple-choice setting, where the output is a single
 130 selected option. The **QA** format serves as an open-ended baseline, requiring the model to produce a
 131 short free-text answer rather than choosing from predefined options. The **List** format generalizes QA
 132 by allowing multiple answers within a single response, capturing scenarios involving alternatives or
 133 multiple plausible options. QA is a special case of the List format, corresponding to a single-element
 134 list. In the List setting, we expect the model to output an ordered list of short textual items, each
 135 representing a potential answer. Because it better reflects clinical uncertainty and our target task
 136 setting, the List format is the primary focus of our experiments. Examples of QA and List outputs
 137 are provided in Figure 1.
 138

139 2.2 EVALUATION

140 **Datasets** For MCQ answer format, we evaluate approaches on standard medical MCQ benchmarks:
 141 (1) **MedQA** (Jin et al., 2021), (2) **MedMCQA** (Pal et al., 2022), (3) **MedXpertQA** (Zuo
 142 et al., 2025) (text), and (4) **MMLU Pro** (Wang et al., 2024) (health). For QA and List answer
 143 formats, we convert the MCQ into *open-ended QA benchmarks* with an LLM-based pipeline similarly
 144 to Myrzakhan et al. (2024). The LLM generates an open-ended version of an MCQ question with
 145 the same ground-truth answer if it deems the question convertible. See Appendices B and D.
 146

147 **Metrics** We evaluate model performance using *accuracy* (Acc) and *mean reciprocal rank* (MRR)
 148 (Radov et al., 2002). Accuracy measures whether the model is able to output the correct ground
 149 truth (choice or exact short answer). This applies to MCQ questions (Acc_{MCQ}), QA (Acc_{QA}), and
 150 ranked-list (Acc_{List}) answer formats. For the list answer format, MRR (MRR_{List}) further evaluates
 151 the *position of the correct item* in the list, assigning higher scores when the correct item appears
 152 earlier. MRR provides a proxy for how effectively a model ranks correct items within its list outputs.
 153

154 Since models may produce correct answers that do not exactly match the ground-truth string for
 155 non-MCQ formats, we also utilize LLM-based evaluation variants (Arora et al., 2025): Acc_{QA}^{LLM} ,
 156 Acc_{List}^{LLM} , and MRR_{List}^{LLM} . All metrics are normalized to $[0, 1]$ and reported as percentages. Additional
 157 details are in Appendix E.
 158

159 **Ranked List Analyses** We utilize additional metrics to assess differences in generated lists: (1)
 160 *average list length*—the mean number of items in valid, non-empty, lists (**VLL**), and (2) *average
 161 correct-answer position* (**CP**). These metrics show how models express uncertainty: CP reflects

162 confidence calibration, while VLL signals potential gaming through over-generation or hallucination
 163 through excessively long outputs. they characterize model behavior under different settings. We also
 164 examine response length and training dynamics in Appendices I and J.

166 2.3 MODEL SELECTION

168 We include state-of-the-art (SOTA) MRMs trained with different approaches in this study:
 169 **HuatuoGPT-01**, SFT on synthetic data from an agentic pipeline (Chen et al., 2024); **m1**, SFT on
 170 distilled data from a teacher model (Huang et al., 2025a); and **AlphaMed**, RFT using MCQ-based
 171 verifiable rewards (Liu et al., 2025a). They share the same initial model, **Qwen2.5-7B-Instruct**
 172 (Qwen et al., 2025), which we select as the *primary model* in this study and as our main baseline for
 173 fine-tuning experiments. We also explore the effects of initial models for RFT in Section 4.3.2.

174 We broaden the comparison on SOTA proprietary and open-weight models. For proprietary models,
 175 we include the **Gemini 2.5** family (Flash Lite, Flash, Pro) (Comanici et al., 2025) to assess intra-
 176 family performance, and **GPT-4.1-mini** (OpenAI, 2025) for comparison with Gemini 2.5 Flash. For
 177 open-weight models, we evaluate **MedGemma** (4B and 27B) (Sellergren et al., 2025), with **Gemma**
 178 **3-4B-Instruct** (Team et al., 2025) (the general counterpart), and **OpenThinker3-7B** (Guha et al.,
 179 2025), a reasoning model built on Qwen2.5-7B-Instruct. Within the Qwen family, we also consider
 180 the **3B**, **14B** sizes, and newer **Qwen3-4B** (Yang et al., 2025). Appendix C provides more details.

182 3 PROMPTING

184 We design six prompt templates derived from three answer formats—**MCQ**, **QA**, and **List**—and two
 185 variants: zero-shot and chain-of-thought (**CoT**). These are used for both evaluation and as prior
 186 prompts for RFT, where prior prompt (Taveekitworachai et al., 2025) guides model behavior *during*
 187 *RFT* and can differ from prompt used *during inference*. The CoT variant follows Xie et al. (2025),
 188 with added answer format instructions. Zero-shot variants omit reasoning instructions (<think>
 189 tags) while retaining the core components. Based on this setup, we expect models to perform best
 190 in the MCQ format, worse in the QA format, and improved in the ranked-list format compared to
 191 QA, since lists increase the likelihood of producing a correct answer in a single inference. Table 1
 192 presents the average performance across benchmarks for each model and answer format, including
 193 the CoT variant (non-aggregated results in Appendix N.1). Additional discussion in Appendix I.1.

194 **MCQ** To set a baseline, we first evaluate MCQ performance. As expected, proprietary models
 195 outperform open-weight ones, even without domain-specific training. Most medical models are
 196 competitive to the proprietary models, with the exception of HuatuoGPT-01. HuatuoGPT-01, SFT
 197 on data synthesized via an agentic pipeline, improves over its base model (Qwen2.5 7B Instruct)
 198 but lags behind other MRMs, often ignoring the required format and producing unparsable outputs
 199 (see Figure 28.) In contrast, m1-SFT on teacher-distilled data—performs competitively, though its
 200 reasoning style mirrors the teacher’s verbosity (see Figure 29). This contrast suggests that designed
 201 synthetic data may introduce greater distributional shift than naturally generated data (Li et al.,
 202 2025b).

204 **MCQ vs. QA** Contrary to the expectation, QA accuracy often exceeded MCQ accuracy (in 7 of
 205 15 cases where $\text{Acc}_{\text{QA}}^{\text{LLM}}$ exceeded Acc_{MCQ} by more than 3 pp., including open-weight models such
 206 as Qwen, Gemma 3 4B, MedGemma 4B, and OpenThinker3), suggesting these models can retain
 207 knowledge independent of format. In contrast, SOTA proprietary models (Gemini 2.5 Pro, GPT-4.1-
 208 mini) and MRMs (HuatuoGPT-01, AlphaMed) declined, indicating reliance on answer-format cues.
 209 Overall, smaller open-weight models appear more robust across formats, while stronger proprietary
 210 and medical models depend more on format cues for knowledge recall (Singh et al., 2025).

212 **QA vs. List** Except for Qwen2.5 7B Instruct and OpenThinker3 7B, all models achieve higher
 213 $\text{Acc}_{\text{List}}^{\text{LLM}}$ than $\text{Acc}_{\text{QA}}^{\text{LLM}}$, likely indicating they often **know the correct answer but fail to select the**
 214 **benchmark’s choice**. Although additional studies are required to confirm these causal effects. The
 215 gap between $\text{Acc}_{\text{List}}^{\text{LLM}}$ and $\text{MRR}_{\text{List}}^{\text{LLM}}$, along with low $\text{MRR}_{\text{List}}^{\text{LLM}}$ and CP, supports this finding (see
 Section 5). Still, most models rank the correct answer near the top, showing they can generally

Table 1: Performance of proprietary, open-weight, and medical LLMs under zero-shot and CoT prompting across MCQ, QA, and List benchmarks. Positive and negative CoT effects are colored; **bold** highlights the best overall score or largest change per metric. For ranked list answers, **CP** is the average rank of the correct item and **VLL** the average length of valid, non-empty, lists.

	MCQ		QA		List				CP	VLL
	Acc _{MCQ}	Acc _{QA}	Acc _{QA} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}			
<i>Proprietary Models</i>										
Gemini 2.5 Flash Lite	48.47	12.35	48.69	14.84	53.82	13.34	46.36	1.39	2.86	
+CoT	-22.88	-1.01	-6.27	-6.74	-29.44	-5.62	-23.42	1.17	1.92	
Gemini 2.5 Flash	55.19	11.95	46.10	17.49	62.66	15.51	53.52	1.41	3.00	
+CoT	-19.75	-0.51	-1.69	-9.99	-35.74	-8.59	-29.40	1.30	2.66	
Gemini 2.5 Pro	58.68	10.89	49.20	18.44	68.46	16.44	58.85	1.40	3.41	
+CoT	-0.62	-0.58	-1.12	-1.94	-3.18	-1.96	-3.09	1.41	3.46	
GPT-4.1 Mini	54.72	9.30	47.02	12.78	61.71	11.42	53.82	1.36	3.26	
+CoT	-7.01	+1.38	+1.98	+0.07	+3.06	-0.04	+2.69	1.36	3.72	
<i>Open-weight General Models</i>										
Qwen2.5 3B Instruct	29.62	5.68	35.66	9.18	38.70	7.35	29.55	1.69	3.09	
+CoT	+3.06	+0.21	+4.53	-1.65	-9.25	-0.25	-3.13	1.27	1.80	
Qwen2.5 7B Instruct (our initial model)	13.43	7.90	43.19	10.38	39.96	8.81	33.07	1.45	2.39	
+CoT	+17.81	+0.56	+0.16	+0.66	+8.57	+0.06	+5.44	1.91	185.72	
Qwen2.5 14B Instruct	35.88	10.41	45.33	13.09	52.84	11.20	43.44	1.48	3.05	
+CoT	+0.45	-0.30	+2.47	-0.45	-4.71	+0.07	-0.82	1.30	2.36	
Qwen3 4B Instruct 2507	43.82	10.47	47.22	11.39	53.01	8.96	40.49	1.70	3.96	
+CoT	-2.70	-0.45	-2.54	+1.20	-2.43	+1.91	+1.56	1.50	3.56	
Gemma 3 4B IT	30.43	6.22	36.62	11.77	47.77	8.72	34.84	1.83	4.68	
+CoT	-2.29	+0.53	+3.54	-2.48	-10.34	-1.53	-5.75	1.67	4.13	
OpenThinker3 7B	27.57	5.27	31.03	5.44	28.56	5.00	24.51	1.43	3.06	
+CoT	-23.82	-0.41	-0.33	-5.32	-27.72	-4.89	-23.77	1.29	4.76	
<i>Open-weight Medical Models</i>										
MedGemma 4B IT	37.09	7.85	43.19	13.34	53.34	9.58	38.65	2.15	95.06	
+CoT	-6.57	+0.42	-0.35	+0.22	-3.50	+0.25	-2.98	3.14	482.76	
MedGemma 27B IT	48.97	12.78	47.64	14.45	50.74	12.78	43.33	1.46	3.26	
+CoT	-16.24	-2.18	-6.68	+0.73	+2.42	+0.74	+2.57	1.43	3.88	
HuatuoGPT o1 7B	17.75	1.10	3.83	10.13	35.68	7.87	27.58	1.70	4.39	
+CoT	-7.54	-0.53	-1.63	-9.84	-35.13	-7.61	-27.10	1.46	2.45	
m1 7B 23K	39.26	7.65	38.97	11.04	50.02	8.26	36.62	2.01	13.39	
+CoT	-7.88	+0.06	-1.71	-5.34	-14.00	-3.64	-7.28	1.64	19.03	
AlphaMed 7B Instruct RL	40.51	0.82	9.46	4.48	19.25	3.27	14.26	1.83	2.59	
+CoT	-3.04	+4.88	+15.00	-0.13	+1.56	+0.20	+2.14	1.89	55.74	
<i>Our Knowledge-Distilled MRMs (based on Qwen2.5 7B Instruct)</i>										
SFT-MCQ	39.60	11.08	48.04	15.48	57.65	11.17	39.71	2.83	141.72	
+CoT	+1.94	+0.64	-0.48	-3.98	-21.80	+0.17	-5.04	1.09	1.46	
SFT-QA	37.67	11.25	46.57	12.94	51.18	10.23	38.92	1.81	13.85	
+CoT	+0.13	+0.01	+0.47	-5.06	-26.39	-2.57	-15.24	1.15	1.78	
SFT-List	10.68	0.24	1.15	15.26	48.91	13.49	41.85	1.41	2.52	
+CoT	+1.94	+0.18	+0.65	-0.30	-0.01	-0.21	-0.31	1.42	2.57	
<i>Our RFT MRMs (based on Qwen2.5 7B Instruct)</i>										
RFT-MCQ	39.34	9.96	46.33	9.89	40.06	8.16	33.00	1.45	2.29	
RFT-QA	36.80	1.04	25.22	0.62	3.59	0.48	2.82	1.67	3.01	
RFT-List-Acc	22.40	4.28	19.01	22.11	56.61	16.17	40.26	2.07	5.94	
RFT-List-MRR	18.23	4.43	21.90	20.96	61.60	15.83	44.89	2.11	16.97	
RFT-List-Judge-MRR	20.49	6.49	30.36	14.86	60.90	12.16	48.68	1.64	4.46	

order plausible options. Lists are usually short (2–5 items), suggesting models generate mostly relevant candidates, though MedGemma 4B is an exception with VLL = 95.06.

Medical fine-tuning improves MedGemma 4B and m1 over their initial models but yields smaller gains for HuatuoGPT-o1 and AlphaMed. HuatuoGPT-o1 underperforms its base model, likely due to overfitting to response format, as previously discussed. In contrast to HuatuoGPT-o1, m1 performs well across answer formats (revisit in Section 4.1), despite also using SFT—the key difference lies in the data source, as previously discussed. However, AlphaMed, an MRM trained with RFT for MCQ, fails to generalize to other answer formats; we revisit this finding again in Section 4.2.

Reassessing CoT with modern models We reassess CoT as a precursor technique to modern reasoning models and find limited benefits. Only Qwen2.5 7B Instruct shows consistent gains, and AlphaMed improves slightly on QA. In most other cases—especially MCQ tasks and the Gemini 2.5 family—CoT provides little or even harms performance. Although CoT generally reduces accuracy, it reliably alters behavior: nearly all models produce longer outputs (except Gemma 3 4B on QA/list

270 tasks), yet this verbosity shows a weak negative correlation with accuracy (see Appendix I.1). CoT
 271 can sometimes improve ranking, as CP decreases or remains stable. This suggests that when models
 272 already know the correct answer, CoT makes them more confident in ranking it higher. The only
 273 exceptions are Qwen2.5 7B Instruct and MedGemma 4B, which increase CP while generating long
 274 lists with repetitive items (185+ and 480+ items; see Section 4.3.2 and appendix L).

277 4 FINE-TUNING

279 For the fine-tuning datasets, we adopt the training set from AlphaMed (Liu et al., 2025a), which
 280 includes MedQA and portions of the MedMCQA training splits. This makes both MedQA and
 281 MedMCQA in-domain benchmarks for evaluation. The MCQ version is used to train MCQ answer
 282 formats, while the open-ended version (see Section 2.2) is used to train QA and list answer formats,
 283 each paired with its respective prior prompt. All experiments are done on 4xH100 GPUs.

285 4.1 SUPERVISED FINE-TUNING (SFT)

287 **Experimental Setup** To obtain a training set, we distill from **Qwen3-30B-A3B-Thinking-2507**
 288 (Yang et al., 2025) (a reasoning variant) on the AlphaMed datasets, prepended with each prompt
 289 template and filtered by rejection sampling (`max_retries=20`; only correct responses retained).
 290 Additional details, including a number of records, are provided in Appendix G. We then fine-
 291 tune **Qwen2.5-7B-Instruct** on the corresponding datasets to obtain three models (SFT-MCQ,
 292 SFT-QA, and SFT-List), one for each answer format considered. Fine-tuning is performed with
 293 LLaMA-Factory (Zheng et al., 2024) using the hyperparameters provided in Appendix H.1.

295 **Results and Discussions** Table 1 shows that SFT on distilled datasets, regardless of answer format,
 296 generally outperforms initial models. Examples of generated responses from each model is available
 297 in Appendix O.1. CoT, however, offers limited benefit to SFT models, similar to what is observed
 298 with m1, regardless of format. CoT slightly improves MCQ performance and CP but often shortens
 299 lists, unlike m1, except in SFT-List. We speculate that, having been trained on a ranked-list format,
 300 CoT leads SFT-List to “overthink,” adding more items to the list to be safe.

301 Given that SFT-MCQ and m1 are trained with a similar approach, differing only in training sets
 302 and hyperparameters, it is unsurprising that SFT-MCQ performs well across formats, similar to
 303 m1. These models achieve the best results not only on MCQ but also on QA (Acc_{QA}^{LLM}) and list
 304 formats (Acc_{List}^{LLM}). However, SFT-MCQ produces unusually long lists ($VLL = 141.72$, $CP = 2.83$,
 305 see Figure 33), and unlike m1, its lists shorten considerably when CoT is applied. The model
 306 initially produces diverse items but soon repeats content, possibly because list-style outputs are
 307 underrepresented in its training data, lowering the likelihood of generating a stop token.

309 While SFT-QA generalizes across formats, though not best on QA, SFT-List achieves the highest
 310 MRR_{List}^{LLM} but performs poorly on Acc_{List}^{LLM} , MCQ, and QA, showing weak robustness across formats.
 311 SFT-List persistently outputs lists—even when instructed with other formats—yielding unextractable
 312 answers. We conjecture that list formats are underrepresented in LLM training, leading to over-
 313 association between list format and knowledge expression.

315 4.2 REINFORCEMENT FINE-TUNING (RFT)

317 RFT (Zhang et al., 2025a) is based on RL using verifiable rewards (Lambert et al., 2025), which can
 318 be directly checked against ground truth, often without a learned reward model. Unlike SFT, where a
 319 model learns fixed input–output pairs, RFT incentivizes responses that maximize reward, balancing
 320 exploration and exploitation. For example, AlphaMed uses RFT to encourage extended reasoning
 321 traces that improve medical MCQ performance without explicit response examples. In this study,
 322 we extend RFT beyond MCQ to the ranked-list format. As reward functions for this format are not
 323 well studied, we design one for our setting and assess its effectiveness. Section 4.3 details ablations
 on the effects of initial models, reward functions, and hyperparameters.

324
325

326 We RFT **Qwen2.5 7B Instruct** using the AlphaMed dataset for the MCQ answer format and QA-
 327 AlphaMed for the QA and ranked-list formats, resulting in five models for the main experiments, as
 328 the ranked-list format includes three types of reward functions (RFT-MCQ, RFT-QA, RFT-List-Acc,
 329 RFT-List-MRR, and RFT-List-Judge-MRR). Training is performed with `ver1` (Sheng et al., 2025),
 330 with hyperparameters and additional details provided in Appendix H.2.

331

332 **Reward Function** The reward function is a crucial component of the RFT setup. It provides
 333 the signal that incentivizes desired model behaviors during training, e.g., assigning high scores to
 334 correct answers. We adopt a design similar to Guo et al. (2025), consisting of two components: (i)
 335 a **correctness reward**, which measures whether the output matches the ground truth, and (ii) an
 336 optional **format reward**, which checks proper output formatting when using CoT prior prompts.
 337 All rewards are normalized to $[0, 1]$, with each component weighted equally.

338

339 **CORRECTNESS REWARD** The correctness reward is task-dependent. Let \hat{y} denote the model
 340 prediction, y^* the ground truth, $\mathbb{1}[\cdot]$ the indicator function, and $N(\cdot)$ a normalization function.
 341 For multiple-choice and QA tasks, we define $R_{\text{MCQ}} = \mathbb{1}[\hat{y} = y^*]$ and $R_{\text{QA}} =$
 342 $\mathbb{1}[N(y^*) \text{ is a substring of } N(\hat{y})]$. For list-style answers, y^* is compared against all items $\hat{Y} =$
 343 $\{\hat{y}_1, \dots, \hat{y}_n\}$, with $R_{\text{List}} = \max_{i=1, \dots, n} R_{\text{QA}}(y^*, \hat{y}_i)$. This treats the list as an *unordered set*: the
 344 model receives full reward if the ground truth appears anywhere in the list, regardless of position.

345

346 **RANK-AWARE CORRECTNESS FOR LISTS** We design a new reward inspired by MRR that in-
 347 centivizes *higher placement* of correct answers. Let r denote the position of the first correct item
 348 (1-indexed). We define $R_{\text{MRR}} = R_{\text{List}}(y^*, \hat{y}_i) \cdot 1/r$. However, models may exploit this by generating
 349 excessively long lists to maximize coverage. A length-penalized variant is discussed in Appendix L.

350

351 **LLM JUDGE REWARD** Inspired by recent studies that use LLM judges as rewards in RFT (Guo
 352 et al., 2025; Su et al., 2025; Zhou et al., 2025), we extend MRR with an LLM-based judge (Judge-
 353 MRR), where semantic equivalence between each \hat{y}_i and y^* is determined by an LLM rather than by
 354 normalized exact match. This provides a more accurate reward signal in cases with varied wording.
 355 We ablate different models and prompts for Judge-MRR in Section 4.3.1.

356

357 **FORMAT REWARD** When using a CoT variant of the prior prompts, we add a format reward (Guo
 358 et al., 2025) that checks whether the output contains exactly one pair of `<think>` and `</think>`
 359 tags, with `<think>` appearing at the very start. The reward is $R_{\text{format}} = 1$ if the conditions are
 360 satisfied, and 0 otherwise. We compare setups with and without this format reward in Section 4.3.1.

361

4.2.2 RESULTS AND DISCUSSIONS

362

363 **General Performance Trend** From Table 1, RFT-MCQ matches SFT-MCQ and m1, confirming
 364 that medical training improves performance over initial models. Like SFT-MCQ, RFT-MCQ per-
 365 forms best on QA benchmarks, while RFT-QA improves but still lags behind. Since MCQ training
 366 benefits both MCQ and QA formats, and list training benefits list tasks, we further examine the
 367 effects of mixed-format training in Appendix K. Most RFT models generate short lists, except RFT-
 368 List-MRR, which—unlike the repetitive outputs of the SFT models—produces long but diverse lists.

369

370 RFT-MCQ and AlphaMed, trained on the same dataset, show similar overall trends but diverge on
 371 QA and list tasks. This may be due to differences in training setup: AlphaMed was trained with
 372 roughly twice as many steps, while RFT-MCQ also incorporates a format component in the reward
 373 function. We analyze the effects of training duration and reward design in Section 4.3.1. We provide
 374 examples of generated responses from the RFT models in Appendix O.2.

375

376 **RFT-QA Exhibits List-like Behavior** We find many cases where RFT-QA generates multiple
 377 answers inside `\boxed{}` during QA evaluation (see Figure 35). This list-like behavior suggests
 378 that under our settings, the model finds it natural to output multiple answers to increase its chance
 379 of receiving a reward, given that the reward function is based on substring-normalized matching.
 380 This constitutes reward exploitation, i.e., reward hacking. However, while such responses achieve
 381 higher rewards during training, they lead to poor evaluation scores on LLM-based metrics, as judges

378 are instructed to penalize this behavior. We also observe that this style carries over to list-format
 379 evaluation, where individual items often contain multiple answers, resulting in significantly lower
 380 scores. This emphasizes the importance of reward design, which we discuss further in Section 4.3.1.
 381

382 **Reward Function Effects With RFT-List** RFT-List models excel on list-based metrics: RFT-
 383 List-MRR achieves the highest Acc_{List}^{LLM} , while RFT-List-Judge-MRR leads on MRR_{List}^{LLM} . All outper-
 384 form SFT-List and, unlike it, also perform robustly on non-list formats. Examples of generated re-
 385 sponses from RFT-List-Judge-MRR under MCQ, QA, and List evaluations are shown in Figures 38,
 386 40 and 41, respectively. This indicates that RFT is not only more robust across answer formats
 387 than SFT, but also extends prior findings that RFT generalizes better to *unseen examples* (Chu et al.,
 388 2025) to the setting of *unseen answer formats*.

389 We conjecture that the same mechanism underlying RFT’s advantage on unseen examples also ap-
 390 plies here: RFT tends to prune irrelevant or inaccurate knowledge from its reasoning paths, reducing
 391 entanglement between knowledge recall and answer format. In contrast, SFT attempts to imitate the
 392 full reasoning trajectory, which can inadvertently bind answer formats to the specific reasoning or
 393 knowledge patterns observed during training (Chu et al., 2025).

394 In addition, reward choice shapes final performance. Among the three list rewards, Acc_{List} boosts
 395 exact matches, while MRR_{List}^{LLM} encourages longer lists, increasing the chance of including the correct
 396 answer and yielding the highest Acc_{List}^{LLM} . The MRR_{List}^{LLM} reward, as expected, produces the best
 397 MRR_{List}^{LLM} . Since none of the current rewards explicitly control list length, MRR_{List} rewards result in
 398 the longest outputs. Appendix L discusses a length-penalized variant of the reward functions.
 399

400 4.3 ABLATION STUDIES FOR RFT

401 In this section, we present ablation studies on factors influencing medical reasoning models trained
 402 with RFT. Section 4.3.1 examines the format component of the reward function, extended training,
 403 and the Judge-MRR setup. Finally, Section 4.3.2 studies the effect of the initial model choice.
 404

405 4.3.1 FACTORS AFFECTING RFT

406 We conduct scaled-down ablation studies to better understand effects of key factors in RFT training,
 407 including the **format component of the reward function**, **extended training duration**, the **role of**
 408 **prior prompts**, and the **choice of judge model**. The full results are available at Appendix I.5.1.
 409

410 **Excluding the format reward or extending training has limited effect on the model** Removing
 411 the format reward does not substantially affect final model performance or list behaviors. Similarly,
 412 extending training from two to four epochs does not substantially improve performance. For ex-
 413 ample, in RFT-MCQ, Acc_{MCQ} remains around 39% with or without the format reward (39.34% →
 414 39.56%), and four epochs very slightly increase it to 39.97%.

416 **Effects of prior prompts** Prior prompts can influence the initial optimization space during RFT
 417 and interact with reward function components. Removing or modifying prior prompts shows mixed
 418 effects, and our ablation scale remains insufficient for conclusive findings (see Appendix I.5.1).
 419

420 **Different judge models** Changing the judge model substantially impacts performance. Replacing
 421 GPT-4.1-mini with Gemini 2.5 Flash improves Acc_{QA}^{LLM} from 30.36% to 43.16% and boosts MCQ
 422 accuracy (20.49% → 33.11%), while maintaining comparable list accuracy (60.90% vs. 59.34%).
 423 In contrast, simplifying the judge prompt severely degrades ranked-list evaluation, with MRR_{List}^{LLM}
 424 falling from 48.68% to 26.19%. This degradation arises because the model exploits weaknesses in
 425 the simplified judge by producing vague or grouped answers, which yield higher training rewards
 426 (see Figure 18a) but fail to generalize at evaluation time. These findings highlight that both the
 427 choice of judge and the design of the judge prompt are critical to final performance.
 428

429 4.3.2 EFFECTS OF INITIAL MODELS USED IN RFT

430 To examine how findings generalize across base models, we extend our RFT setup to **smaller mod-
 431 els** (*Qwen2.5 3B Instruct*), **more recent model families** (*Qwen3 4B Instruct*), and **continual RFT**

432 **from existing reasoning models.** We consider three scenarios: continual RFT from (1) a **general**
 433 **reasoning model**, *OpenThinker3* (domain adaptation); (2) **an MRM trained with SFT-MCQ**, *m1*;
 434 and (3) **an MRM trained with RFT-MCQ**, *AlphaMed*. Appendix I.5.2 provides full results.
 435

436 **Model family and scale** Qwen3 4B after RFT becomes competitive with Gemini 2.5 Pro (RFT-
 437 List-Acc’s 53.01% → 71.60% vs. 68.46% $\text{Acc}_{\text{List}}^{\text{LLM}}$, and RFT-List-MRR’s 47.22% → 48.54% vs.
 438 49.20% $\text{Acc}_{\text{List}}^{\text{LLM}}$). Despite being similar in size to Qwen2.5 3B, Qwen3 4B is consistently stronger.
 439 Qwen2.5 3B struggles with RFT-QA setup (27.60%) but benefits from RFT-List setups (35.66% →
 440 40.20% $\text{Acc}_{\text{QA}}^{\text{LLM}}$, 38.70% → 59.82% $\text{Acc}_{\text{List}}^{\text{LLM}}$). We conjecture that RFT-List setups provide denser
 441 signals that transfer to QA, as the model can attempt multiple answers in a single inference call,
 442 allowing it to incorporate more from the training data.
 443

444 **Continual RFT benefits reasoning models** For OpenThinker3, MCQ and QA performance
 445 converge to a similar range across RFT setups, but the RFT-List setup yields a substantial boost: Acc_{MCQ}
 446 rises from 27.57% to 33.74–34.60%, and $\text{Acc}_{\text{QA}}^{\text{LLM}}$ from 31.03% to 39.78–41.42%. In contrast, for
 447 list evaluations, non-list RFT models reach only 33.28–39.02% $\text{Acc}_{\text{List}}^{\text{LLM}}$, whereas list-based RFT
 448 jumps to 56.98–59.44%, with $\text{MRR}_{\text{List}}^{\text{LLM}}$ improving from 24.51% to 35.03%.
 449

450 A similar pattern holds for *m1*; all RFT setups bring MCQ and QA into a comparable range. How-
 451 ever, only RFT-List improve list-format accuracy, while RFT-MCQ and RFT-QA reduce the per-
 452 formance. These results suggest that SFT → RFT is most beneficial when the setup is RFT-List.
 453

454 AlphaMed further illustrates the benefits of sequencing: after initial RFT-MCQ, subsequent RFT-
 455 QA lifts $\text{Acc}_{\text{QA}}^{\text{LLM}}$ from 9.46% → 38.35% and RFT-List-Acc training improves $\text{Acc}_{\text{List}}^{\text{LLM}}$ from 19.25%
 456 → 57.29%, while retaining MCQ ability. However, RFT-MCQ→RFT-MCQ degrades $\text{Acc}_{\text{List}}^{\text{LLM}}$, re-
 457 inforcing that MCQ is easier but less transferable than QA/List answer formats.
 458

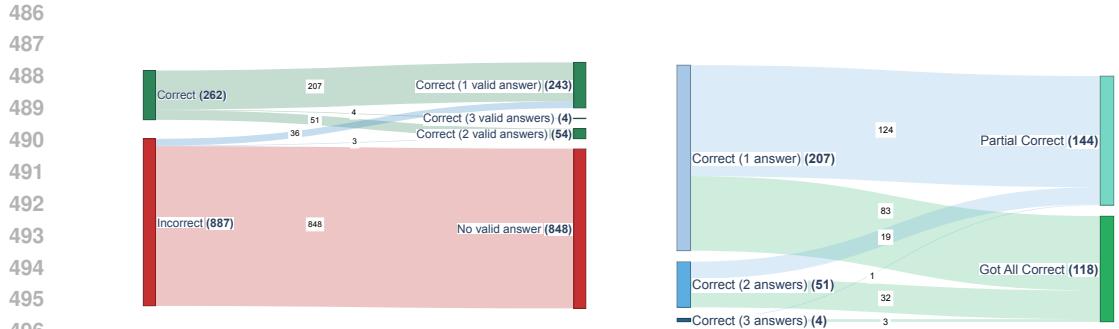
459 **List rewards may incentivize a large list** RFT-List setups often produce excessively long lists;
 460 for instance, Qwen2.5 3B, Qwen3 4B, OpenThinker3, and *m1* average over 700 items under RFT-
 461 List-Acc/MRR training. This may stem from repetition at the tail end and a reduced probability
 462 of generating a stop token. AlphaMed does not show this behavior, suggesting that initial MCQ
 463 training stabilizes later QA/List training. This supports the view that curriculum learning in RFT
 464 is beneficial not only for sequencing data difficulty (Stojanovski et al., 2025; Xie et al., 2025), but
 465 also for sequencing answer formats—from simpler MCQ to more complex QA or list outputs. See
 466 Appendix K for further discussion and Appendix L for length-penalized rewards.
 467

4.4 DISCUSSIONS

468 The reward functions introduced in Section 4.2.1 improve MRM accuracy across multiple answer
 469 formats and represent a step toward enhancing ranked-list generation. While these rewards implicitly
 470 convey a notion of model confidence—e.g., correct items should rank higher—they do not explicitly
 471 quantify the differences in confidence magnitudes. Knowing how strongly the model’s confidence
 472 in rank 1 over rank 2 is valuable in some application, but supporting such calibrated confidence
 473 would require an additional training objective. Our approach is orthogonal to confidence calibration
 474 in RFT settings. For example, Damani et al. (2025) introduce calibration-oriented rewards, which
 475 are complementary to our reward functions.
 476

477 Another consideration is that our experiments focus on academic benchmark settings, where per-
 478 formance (e.g., accuracy) is the primary metric. Additional objectives and safeguards would be
 479 required to translate our work to real-world clinical use. For instance, a fine-tuned model optimizing
 480 for benchmark accuracy may overlook practical constraints—such as avoiding potentially harmful
 481 suggestions or ensuring the inclusion of rare but clinically important conditions. These challenges
 482 reflect limitations in existing datasets, which often provide only a single ground-truth answer rather
 483 than multiple valid options. Future work could address this by constructing more realistic bench-
 484 marks or introducing reward functions that explicitly account for safety and clinical relevance.
 485

486 Overall, our study aims to provide a foundation and a new perspective on answer-format-aware
 487 training for MRMs. Rather than replacing clinical experts, we view this direction as supporting more
 488 effective collaboration between models and practitioners. We hope this work encourages further
 489



(a) Records shifting between *correct* and *incorrect* from single- to multi-answer evaluation. (b) Correct responses split into those covering all vs. partial valid answers under multi-answer evaluation.

Figure 2: Re-evaluation of MedQA with multiple valid answers.

research into principled, clinically grounded approaches for training MRMs to generate richer and safer answer formats.

5 CASE STUDY: MEDQA WITH MULTIPLE VALID ANSWERS

As uncertainty—such as incomplete patient information—is common in real-world settings, multiple answers may be valid. A ranked-list format can broaden clinical perspectives and mitigate cognitive blind spots, inviting diverse views to guide patient care. However, these constraints are not fully accounted for in current medical benchmarks, which mostly rely on single-answer evaluation.

To demonstrate this, we adapt MedQA by adding metadata with multiple expert-annotated valid answers from Saab et al. (2024), creating a modified version where each record includes several valid answers. We then take ranked lists generated by **RFT-List-Judge-MRR** (our best model on MRR_{List}^{LLM}), from its original QA-MedQA evaluation, and re-evaluate them on the modified benchmark using a normalized exact-match comparison.

Results from Figure 2a show that, out of 1,149 records, 43 were classified as *incorrect* under single-answer evaluation but actually contained *valid answers* according to the modified benchmark. This suggests that the model may already possess sufficient knowledge to generate valid responses; however, it fails to select the answer that matches the benchmark’s preferred label, highlighting inherent biases in benchmark development.

Furthermore, Figure 2b shows that 133 of the original 233 correct answers—including 55 records with more than one valid answer—contained all expert-validated answers within the generated lists. This indicates that models can capture nuanced sets of valid options when generating lists. Nevertheless, the current model is still limited: not all generated list answers include all valid cases. This limitation is unsurprising given that most training datasets and benchmarks nowadays are single-best-answer formats. Addressing this gap through improved datasets that capture the nuances of medical applications represents a promising future direction beyond the scope of this work.

6 CONCLUSIONS

We study how prompting and fine-tuning steer MRMs toward generating ranked-list answers. RFT generalizes better than SFT across diverse answer formats, and ranking-oriented rewards (e.g., MRR, Judge-MRR) further improve performance under ranked-list evaluations. Our study provides the first systematic comparison of prompting vs. fine-tuning for ranked-list generation in medicine, introduces reward functions inspired by MRR, and includes comprehensive ablations analyzing how initial models, reward design, and hyperparameters shape RFT performance. We hope this work motivates development of models and benchmarks that better reflect real-world clinical practice, moving beyond the limitations of single-answer evaluation.

540
541
LIMITATIONS542
543
544
545
546
547
548
549
550
As one of the first studies to investigate ranked-list answer generation for medical reasoning models, our work opens several promising directions for future research. In this paper, we focus exclusively on ranked lists, but alternative non-list formats could be explored in subsequent studies. Although our evaluation is limited to the medical domain, the underlying approach may generalize to other fields. Further progress will also require larger-scale experimentation. We also restrict our study to a monolingual and text-only setup. Extending the approach to multimodal and multilingual settings is an exciting avenue for future research. Given our limited compute budget, these results are informative in our setting but may not fully generalize to larger-scale RFT training, e.g., larger models, longer training, and larger datasets; however, we expect several findings to hold.551
552
553
554
555
While we report behavioral metrics, we do not conduct a detailed analysis of reasoning traces, as our emphasis is on the final answer format. Changing this format may itself influence reasoning traces, which we leave for future investigation. Similarly, more extensive exploration of data, algorithms, and hyperparameters—beyond what was feasible under our computational budget—remains an important direction.556
557
558
559
560
561
562
563
Another limitation is the lack of medical benchmarks with multiple correct or ranked answers. To approximate this setting, we adapted multiple-choice QA datasets into a ranked-answer format. Developing benchmarks that better capture the diversity of real-world clinical reasoning is beyond the scope of this work but represents a valuable opportunity for the community. In addition, we focus on a single-turn setting, consistent with traditional accuracy-oriented benchmarks. Evaluating and developing models that can respond in multiple formats within multi-turn interactions—such as sequential diagnosis, where ranked lists may be updated as new information becomes available, or dynamic settings like simulated patient encounters—is an interesting direction for future work.564
565
566
567
568
Finally, ranked lists themselves have limitations. While they indicate which answers a model considers most likely, they do not reflect differences in probability magnitude. For instance, the top-ranked item may have a 50% likelihood, while the rest trail far behind. Capturing such disparities is an open challenge. We speculate that orthogonal approaches, such as those explored by Stangel et al. (2025); Damani et al. (2025), could complement ranked-list methods.569
570
ETHICS STATEMENT
571572
573
574
575
576
This work evaluates methods for generating ranked medical answers using reasoning models, tested on academic benchmarks that may not fully reflect real-world clinical environments—an acknowledged limitation that our methods aim to partly address. LLMs remain prone to hallucinations, and their outputs must be interpreted with caution; they are not substitutes for professional judgments.577
578
REPRODUCIBILITY STATEMENT
579580
581
582
583
We are committed to ensuring the reproducibility of our work. Detailed descriptions of models, datasets, training procedures, and evaluation protocols are provided in Section 2. Additional implementation details and training hyperparameters are available throughout the Appendix. In addition, datasets and source code are publicly available at the links in Appendix B and Appendix H.2.584
585
REFERENCES
586587
588
589
590
Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025. URL <https://arxiv.org/abs/2505.08775>.591
592
Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-01, towards medical complex reasoning with llms, 2024. URL <https://arxiv.org/abs/2412.18925>.

594 Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V
 595 Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of founda-
 596 tion model post-training. In *Forty-second International Conference on Machine Learning*, 2025.
 597 URL <https://openreview.net/forum?id=dYur3yabMj>.

598 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit
 599 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marrs, Sam Petulla, Colin
 600 Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-
 601 Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric
 602 Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, et al.
 603 Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next
 604 generation agentic capabilities, 2025. URL <https://arxiv.org/abs/2507.06261>.

605 Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob
 606 Andreas. Beyond binary rewards: Training lms to reason about their uncertainty, 2025. URL
 607 <https://arxiv.org/abs/2507.16806>.

608 Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. AI
 609 hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In
 610 Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and
 611 Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational
 612 Linguistics*, pp. 10183–10213, Abu Dhabi, UAE, January 2025. Association for Computational
 613 Linguistics. URL <https://aclanthology.org/2025.coling-main.680/>.

614

615 Tingchen Fu, Jiawei Gu, Yafu Li, Xiaoye Qu, and Yu Cheng. Scaling reasoning, losing control:
 616 Evaluating instruction following in large reasoning models, 2025. URL <https://arxiv.org/abs/2505.14810>.

617

618 Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: a systematic review of
 619 frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.*, 19(1):121–127, January
 620 2012.

621

622 Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. Large language models
 623 lack essential metacognition for reliable medical reasoning. *Nature Communications*, 16(1):642,
 624 Jan 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55628-6. URL <https://doi.org/10.1038/s41467-024-55628-6>.

625

626 Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna
 627 Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu
 628 Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su,
 629 Wanja Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan
 630 Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak,
 631 Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia
 632 Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill,
 633 Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy,
 634 Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models,
 635 2025. URL <https://arxiv.org/abs/2506.04178>.

636

637 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu
 638 Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhi-
 639 hong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng,
 640 Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie
 641 Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei
 642 Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Ji-
 643 ashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi
 644 Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin
 645 Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu,
 646 Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun
 647 Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen,
 648 Qishi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi
 649 Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping
 650 Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian

648 Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin
 649 Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xi-
 650 aotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng
 651 Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaox-
 652 iang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang,
 653 Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu,
 654 Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan,
 655 Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou,
 656 Yuji He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu,
 657 Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yut-
 658 ing Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang,
 659 Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li,
 660 Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen
 661 Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*,
 662 645(8081):633–638, Sep 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL
 663 <https://doi.org/10.1038/s41586-025-09422-z>.

664 Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. m1: Unleash the potential
 665 of test-time scaling for medical reasoning with large language models, 2025a. URL <https://arxiv.org/abs/2504.00869>.

666 Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei
 667 Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey – part 2: Surpassing o1-preview
 668 through simple distillation, big progress or bitter lesson?, 2024. URL <https://arxiv.org/abs/2411.16489>.

669 Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei
 670 Liu, and Xiaofan Zhang. O1 replication journey – part 3: Inference-time scaling for medical
 671 reasoning, 2025b. URL <https://arxiv.org/abs/2501.06458>.

672 Sarah Jabbour, David Fouhey, Stephanie Shepard, Thomas S. Valley, Ella A. Kazerooni, Nikola
 673 Banovic, Jenna Wiens, and Michael W. Sjoding. Measuring the impact of ai in the diagnosis of
 674 hospitalized patients: A randomized clinical vignette survey study. *JAMA*, 330(23):2275–2284,
 675 12 2023. ISSN 0098-7484. doi: 10.1001/jama.2023.22295. URL <https://doi.org/10.1001/jama.2023.22295>.

676 Doohyuk Jang, Yoonjeon Kim, Chanjae Park, Hyun Ryu, and Eunho Yang. Reasoning model
 677 is stubborn: Diagnosing instruction overriding in reasoning models, 2025. URL <https://arxiv.org/abs/2505.17225>.

678 Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What dis-
 679 ease does this patient have? a large-scale open domain question answering dataset from medical
 680 exams. *Applied Sciences*, 11(14), 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL
 681 <https://www.mdpi.com/2076-3417/11/14/6421>.

682 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph
 683 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language
 684 model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Sys-
 685 tems Principles*, SOSP ’23, pp. 611–626, New York, NY, USA, 2023. Association for Com-
 686 puting Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL <https://doi.org/10.1145/3600006.3613165>.

687 Deepti Lamba, William H. Hsu, and Majed Alsadhan. Chapter 1 - predictive analytics and ma-
 688 chine learning for medical informatics: A survey of tasks and techniques. In Pardeep Ku-
 689 mar, Yugal Kumar, and Mohamed A. Tawhid (eds.), *Machine Learning, Big Data, and IoT for
 690 Medical Informatics*, Intelligent Data-Centric Systems, pp. 1–35. Academic Press, 2021. ISBN
 691 978-0-12-821777-1. doi: <https://doi.org/10.1016/B978-0-12-821777-1.00023-9>. URL <https://www.sciencedirect.com/science/article/pii/B9780128217771000239>.

692 701 Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brah-
 693 man, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya

702 Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christo-
 703 pher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Ha-
 704 jishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on*
 705 *Language Modeling*, 2025. URL <https://openreview.net/forum?id=iluGbfHHpH>.

706 Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice
 707 questions really be useful in detecting the abilities of LLMs? In Nicoletta Calzolari, Min-Yen
 708 Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of*
 709 *the 2024 Joint International Conference on Computational Linguistics, Language Resources and*
 710 *Evaluation (LREC-COLING 2024)*, pp. 2819–2834, Torino, Italia, May 2024. ELRA and ICCL.
 711 URL <https://aclanthology.org/2024.lrec-main.251/>.

712 Xiaomin Li, Zhou Yu, Zhiwei Zhang, Xupeng Chen, Ziji Zhang, Yingying Zhuang, Narayanan
 713 Sadagopan, and Anurag Beniwal. When thinking fails: The pitfalls of reasoning for instruction-
 714 following in llms, 2025a. URL <https://arxiv.org/abs/2505.11423>.

715 Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ra-
 716 masubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners.
 717 In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.),
 718 *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25366–25394, Vi-
 719 enna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-
 720 5. doi: 10.18653/v1/2025.findings-acl.1301. URL <https://aclanthology.org/2025.findings-acl.1301/>.

721 Seungseop Lim, Gibaeg Kim, Hyunkyoung Lee, Wooseok Han, Jean Seo, Jaehyo Yoo, and Eunho
 722 Yang. H-ddx: A hierarchical evaluation framework for differential diagnosis, 2025. URL
 723 <https://arxiv.org/abs/2510.03700>.

724 Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao
 725 Jiang, Mengze Li, Song xiaohui, Siliang Tang, Jun Xiao, Hui Lin, Yeting Zhuang, and Beng Chin
 726 Ooi. HealthGPT: A medical large vision-language model for unifying comprehension and gen-
 727 eration via heterogeneous knowledge adaptation. In *Forty-second International Conference on*
 728 *Machine Learning*, 2025. URL <https://openreview.net/forum?id=WbP20wMULq>.

729 Che Liu, Haozhe Wang, Jiazen Pan, Zhongwei Wan, Yong Dai, Fangzhen Lin, Wenjia Bai, Daniel
 730 Rueckert, and Rossella Arcucci. Beyond distillation: Pushing the limits of medical llm reasoning
 731 with minimalist rule-based rl, 2025a. URL <https://arxiv.org/abs/2505.17952>.

732 Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and
 733 Min Lin. Understanding rl-zero-like training: A critical perspective, 2025b. URL <https://arxiv.org/abs/2503.20783>.

734 Xiangbin Meng, Xiangyu Yan, Kuo Zhang, Da Liu, Xiaojuan Cui, Yaodong Yang, Muhan Zhang,
 735 Chunxia Cao, Jingjia Wang, Xuliang Wang, Jun Gao, Yuan-Geng-Shuo Wang, Jia-Ming Ji, Zifeng
 736 Qiu, Muzi Li, Cheng Qian, Tianze Guo, Shuangquan Ma, Zeying Wang, Zexuan Guo, Youlan
 737 Lei, Chunli Shao, Wenya Wang, Haojun Fan, and Yi-Da Tang. The application of large language
 738 models in medicine: A scoping review. *iScience*, 27(5):109713, May 2024.

739 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke
 740 Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time
 741 scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.

742 Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From
 743 multi-choice to open-style questions for llms evaluation, benchmark, and arena, 2024. URL
 744 <https://arxiv.org/abs/2406.07545>.

745 Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc
 746 Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, Bay Gross,
 747 Peter Hames, Mustafa Suleyman, Dominic King, and Eric Horvitz. Sequential diagnosis with
 748 language models, 2025. URL <https://arxiv.org/abs/2506.22405>.

749 OpenAI. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>, April
 750 2025. Accessed: 2025-09-08.

756 Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale
 757 multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores,
 758 George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Con-
 759 ference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Re-
 760 search*, pp. 248–260. PMLR, 07–08 Apr 2022. URL <https://proceedings.mlr.press/v174/pal22a.html>.

761

762 Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan,
 763 Hector Liu, Yuanzhi Li, and Pengfei Liu. O1 replication journey: A strategic progress report –
 764 part 1, 2024. URL <https://arxiv.org/abs/2410.18982>.

765

766 Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan
 767 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
 768 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
 769 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,
 770 Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang,
 771 Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.
 772 URL <https://arxiv.org/abs/2412.15115>.

773

774 Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. Evaluating web-based question answer-
 775 ing systems. In Manuel González Rodríguez and Carmen Paz Suárez Araujo (eds.), *Proceedings
 776 of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las
 777 Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).
 778 URL <https://aclanthology.org/L02-1301/>.

779

780 Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations
 781 toward training trillion parameter models, 2020. URL <https://arxiv.org/abs/1910.02054>.

782

783 Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang,
 784 Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike
 785 Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil
 786 Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste
 787 Alayrac, Neil Houldsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai,
 788 Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri,
 789 Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mans-
 790 field, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause,
 791 Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals,
 792 Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Web-
 793 ster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan
 Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine, 2024. URL
<https://arxiv.org/abs/2404.18416>.

794

795 Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha.
 796 A systematic survey of prompt engineering in large language models: Techniques and applica-
 797 tions, 2025. URL <https://arxiv.org/abs/2402.07927>.

798

799 Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si,
 800 Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav
 801 Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao,
 802 Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncenarenco,
 803 Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat,
 804 Alexander Hoyle, and Philip Resnik. The prompt report: A systematic survey of prompt engi-
 neering techniques, 2025. URL <https://arxiv.org/abs/2406.06608>.

805

806 Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo
 807 Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar,
 808 Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy
 809 Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Ken-
 neth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny
 Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak,

810 Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane
 811 Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos,
 812 Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin,
 813 Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard
 814 Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim,
 815 Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Vic-
 816 tor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F.
 817 Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical
 818 report, 2025. URL <https://arxiv.org/abs/2507.05201>.

819 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
 820 Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathe-
 821 matical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.

822 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,
 823 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings
 824 of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297. ACM,
 825 March 2025. doi: 10.1145/3689031.3696075. URL <http://dx.doi.org/10.1145/3689031.3696075>.

826 Shrutiika Singh, Anton Alyakin, Daniel Alexander Alber, Jaden Stryker, Ai Phuong S Tong, Karl
 827 Sangwon, Nicolas Goff, Mathew de la Paz, Miguel Hernandez-Rovira, Ki Yun Park, Eric Claude
 828 Leuthardt, and Eric Karl Oermann. It is too many options: Pitfalls of multiple-choice questions in
 829 generative ai and medical education, 2025. URL <https://arxiv.org/abs/2503.13508>.

830 Paul Stangel, David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Kamilia Zaripova, Matthias Ke-
 831 icher, and Nassir Navab. Rewarding doubt: A reinforcement learning approach to calibrated
 832 confidence expression of large language models, 2025. URL <https://arxiv.org/abs/2503.02623>.

833 Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kad-
 834 dour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning
 835 with verifiable rewards, 2025. URL <https://arxiv.org/abs/2505.24760>.

836 Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu.
 837 Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains, 2025.
 838 URL <https://arxiv.org/abs/2503.23829>.

839 Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu,
 840 Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey
 841 on efficient reasoning for large language models, 2025. URL <https://arxiv.org/abs/2503.16419>.

842 Liyuan Tao, Chen Zhang, Lin Zeng, Shengrong Zhu, Nan Li, Wei Li, Hua Zhang, Yiming Zhao,
 843 Siyan Zhan, and Hong Ji. Accuracy and effects of clinical decision support systems integrated
 844 with bmj best practice–aided diagnosis: Interrupted time series study. *JMIR Med Inform*, 8(1):
 845 e16912, Jan 2020. ISSN 2291-9694. doi: 10.2196/16912. URL <https://doi.org/10.2196/16912>.

846 Pittawat Taveekitworachai, Potsawee Manakul, Sarana Nutanong, and Kunat Pipatanakul. Prior
 847 prompt engineering for reinforcement fine-tuning, 2025. URL <https://arxiv.org/abs/2505.14157>.

848 Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej,
 849 Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas
 850 Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon,
 851 Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai
 852 Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman,
 853 Yi Gao, et al. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.

864 Basil Varkey. Principles of clinical ethics and their application to practice. *Medical Principles and*
 865 *Practice*, 30(1):17–28, 06 2020. ISSN 1011-7571. doi: 10.1159/000509119. URL <https://doi.org/10.1159/000509119>.

867 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=y10DM6R2r3>.

874 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

880 Tong Wu, Chong Xiang, Jiachen T. Wang, G. Edward Suh, and Prateek Mittal. Effectively controlling reasoning models through thinking intervention, 2025. URL <https://arxiv.org/abs/2503.24370>.

884 Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning, 2025. URL <https://arxiv.org/abs/2502.14768>.

887 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

896 Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, Fangfu Liu, Xiang Xu, Jiaze Ma, Xuekai Zhu, Ermo Hua, Yihao Liu, Zonglin Li, Huayu Chen, Xiaoye Qu, Yafu Li, Weize Chen, Zhenzhao Yuan, Junqi Gao, Dong Li, Zhiyuan Ma, Ganqu Cui, Zhiyuan Liu, Biqing Qi, Ning Ding, and Bowen Zhou. A survey of reinforcement learning for large reasoning models, 2025a. URL <https://arxiv.org/abs/2509.08827>.

903 Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning, 2025b. URL <https://arxiv.org/abs/2506.12307>.

907 Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860, August 2023. ISSN 2150-8097. doi: 10.14778/3611540.3611569. URL <https://doi.org/10.14778/3611540.3611569>.

913 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyuan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.38. URL <https://aclanthology.org/2024.acl-demos.38/>.

918 Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Jiale Zhao, Jingwen Yang, Jianwei Lv,
919 Kongcheng Zhang, Yihe Zhou, Hengtong Lu, Wei Chen, Yan Xie, and Mingli Song. Breaking
920 the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning,
921 2025. URL <https://arxiv.org/abs/2508.16949>.

922 Yuxuan Zhou, Xien Liu, Chen Ning, Xiao Zhang, and Ji Wu. Reliable and diverse evaluation of llm
923 medical knowledge mastery, 2024. URL <https://arxiv.org/abs/2409.14302>.

925 Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning
926 Ding, and Bowen Zhou. MedxpertQA: Benchmarking expert-level medical reasoning and un-
927 derstanding. In *Forty-second International Conference on Machine Learning*, 2025. URL
928 <https://openreview.net/forum?id=IyVcxU0RKI>.

929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971

972

973

974

975

976

977

978

Appendix

Table of Contents

979

A Related Work	20
A.1 Medical Reasoning Model	20
A.2 Prompting	20
A.3 Fine-Tuning To Obtain A Reasoning Model	20
A.4 Medical Evaluation	21
B Dataset Overview	21
C Model Overview	22
D Pipeline for Converting MCQs to QA Format	22
E Evaluation Metrics	24
E.1 Additional Details on Metrics	24
E.2 LLM-Based Metrics	24
F Prompt Templates	27
G SFT Training Dataset Preparation	29
H Training Hyperparameters	30
H.1 SFT	30
H.2 RFT	30
I Additional Discussions	31
I.1 Prompting Results	31
I.2 Fine-Tuning Results	32
I.3 SFT Results	32
I.4 RFT Results	32
I.5 Ablation Studies for RFT	32
I.5.1 RFT Factors	32
I.5.2 Initial Models	34
J Training Dynamics	35
J.1 Main Experiment	35
J.2 Factors Affecting RFT	36
J.3 Initial Models	37
K RFT With Mixed Answer Format Datasets	39
L List Reward Hacking and Mitigation With Length Penalty	40
M Length Penalty Hyperparameters	41
N Full Results	43
N.1 Prompting	43
N.2 Fine-Tuning	49
N.2.1 SFT	49
N.2.2 RFT	49
O Qualitative Examples	49

1026	O.1 SFT Models	55
1027	O.2 RFT Models	56
1028		

1029
1030 **LLM USAGE STATEMENT**
1031

1032 LLMs were used only for supportive tasks such as proofreading, grammar refinement, and polishing
1033 text. The research ideas, initial drafts, and core content were developed entirely by the authors;
1034 LLMs did not generate the main body of the manuscript. We also used LLMs to assist with code
1035 snippets for evaluation and analysis, but all codes and logic were reviewed by the authors.
1036

1037 **A RELATED WORK**
1038

1039 **A.1 MEDICAL REASONING MODEL**
1040

1041 The popularity of reasoning models in verifiable domains (e.g. mathematics and coding) (Zhang
1042 et al., 2025a) has inspired several studies in the medical domain to develop a medical reasoning
1043 model with similar approaches. Early work includes HuatuoGPT-01 (Chen et al., 2024), which pro-
1044 pose a data synthesis pipeline for supervised fine-tuning (SFT) and then perform reinforcement fine-
1045 tuning (RFT). Next, m1 (Huang et al., 2025a) investigates knowledge distillation from DeepSeek
1046 R1, an approach similarly to s1 for verifiable domains (Muennighoff et al., 2025). Later, AlphaMed
1047 (Liu et al., 2025a) proposed an approach utilizing RFT only, similarly to DeepSeek-R1-Zero. In
1048 contrast to these approaches, the objective of this paper aimed at investigating settings where rea-
1049 soning models must answer as a list. Recently, Med-U1 (Zhang et al., 2025b) was released as an
1050 attempt to generalize RFT beyond MCQ settings. Med-U1 moves in a similar direction to ours,
1051 though with a different focus. Specifically, Med-U1 explores three answer formats—MCQ, numeric,
1052 and short answer—using different types of rewards for each. By contrast, we examine MCQ, short
1053 answer, and ranked-list answer formats, and we further investigate a variety of reward functions,
1054 including LM-based reward functions.
1055

1056 **A.2 PROMPTING**
1057

1058 A notable example of prompting approaches is chain-of-thought prompting (Wei et al., 2022), which
1059 serves as a precursor to reasoning models. While there is a large body of work on prompting with
1060 LLMs, prompting in reasoning models has been less explored, and we lack a systematic under-
1061 standing of how well prompting approaches that are effective in LLMs transfer to reasoning models.
1062 Furthermore, recent studies have shown that reasoning model has reduced instruction following ca-
1063 pabilities (Li et al., 2025a; Fu et al., 2025; Jang et al., 2025), which directly associated with ability to
1064 prompt an LM. This study aims to systematically investigate whether reasoning models and modern
1065 LLMs can be prompted to generate answers in different formats.
1066

1067 **A.3 FINE-TUNING TO OBTAIN A REASONING MODEL**
1068

1069 There are two main fine-tuning approaches in turning an LLM into a reasoning model: 1) SFT on
1070 responses with reasoning traces and 2) RFT with a verifiable reward function. SFT requires responses
1071 with reasoning traces to fine-tune a model to imitate the style of a reasoning model’s answers. Often,
1072 gathering these training sets are often done by generating responses from existing reasoning models,
1073 i.e., knowledge distillation. In the field of medical reasoning models, m1 utilized this approach.
1074

1075 In contrast, RFT does not teach reasoning to the LM directly, but rather incentivizes the models
1076 to generate intermediate tokens, often called thoughts, that maximize the defined rewards during
1077 training. Reward design is one of the core factors of RFT for incentivization. In MCQ settings, the
1078 reward design is a simple verifiable reward on the correct choices. The core component of reward is
1079 often the accuracy reward, i.e., if the model generates a correct choice matched with the ground truth,
1080 the model scored the reward, and vice versa. Similar approach was done in Liu et al. (2025a) and
1081 Zhang et al. (2025b). To extend beyond MCQ, Zhang et al. (2025b) introduced a different reward
1082 for different scenarios: allowable range for numerical value, exact match and rouge-1 for open-
1083 ended QA. For the ranked list format, our work borrows the fundamental idea from the information
1084

1080 retrieval field-mean reciprocal rank (MRR)—which award higher score if the correct answer is higher
 1081 in rank. We use MRR as reward to incentivize an LM to generate a ranked list answer using the RFT
 1082 approach. We also study incorporating additional terms to penalize long list to investigate effects of
 1083 reward on reasoning model performance and behaviors. In addition, we also experiment with using
 1084 LM judge as a reward.

1085

1086 A.4 MEDICAL EVALUATION

1087

1088 A growing body of work has explored how to evaluate LLMs in medical settings, often focusing on
 1089 dataset construction, multimodal understanding, or complex interactive environments. Zhou et al.
 1090 (2024) examines how to create grounded benchmark datasets leveraging existing medical knowl-
 1091 edge bases. Unlike this line of work, we do not introduce a new dataset; rather, we study robustness
 1092 to answer formats—especially ranked lists—and focus specifically on MRMs, not general-purpose
 1093 medical LLMs. Lin et al. (2025) evaluates models on image-based tasks such as X-ray interpre-
 1094 tation, ultrasound reading, and image reconstruction. These tasks require vision-language capabilities,
 1095 which differ fundamentally from our focus on text-to-text MRMs. As a result, multimodal compre-
 1096 hension lies outside the scope of our work.

1097

1098 Another line of research studies *differential diagnosis* and *hierarchical evaluation*. H-DDx (Lim
 1099 et al., 2025) introduces a pipeline that maps free-text diagnoses to ICD-10 codes and evaluates them
 1100 using hierarchical metrics. Although we reference differential diagnosis as one potential use case
 1101 for ranked lists, our goal differs: we investigate answer-format robustness rather than proposing
 1102 new diagnostic frameworks or hierarchical metrics. Accordingly, our evaluation uses standard rank-
 1103 aware metrics like MRR, which align with single-answer benchmarks.

1104

1105 A recent study has also explored more realistic or interactive evaluation settings. AI Hospital (Fan
 1106 et al., 2025) evaluates LLMs in multi-agent, multi-turn clinical environments to simulate complex
 1107 medical interactions. Similarly, Nori et al. (2025) studies iterative medical reasoning, where a model
 1108 gathers information and revises its diagnosis across multiple turns. In contrast, our study is carried
 1109 out in a single-turn, accuracy-focused setting designed to isolate the effects of answer format and
 1110 training method on MRMs.

1111

1112 Finally, works such as Griot et al. (2025); Arora et al. (2025) evaluate higher-level clinical capa-
 1113 bilities including metacognition and rubric-based reasoning. While we also use LLM judges in our
 1114 evaluation, we follow a reference-based comparison between generated and ground-truth answers
 1115 rather than rubric-based scoring. Overall, our work complements these efforts by offering the first
 1116 systematic examination of how MRMs trained on one answer format generalize when instructed to
 1117 produce another, with particular emphasis on the ranked-list format.

1118

1119 B DATASET OVERVIEW

1120

1121 Table 2 summarizes the datasets used in this study, including both training data and evaluation
 1122 benchmarks. We report the dataset splits, the number of instances, the task format (MCQ or QA),
 1123 and the associated license terms. All datasets are publicly available under licenses that permit their
 1124 use for research purposes.

1125

1126 The QA versions of the datasets are obtained through the conversion pipeline described in Ap-
 1127 pendix D. This process ensures consistency between the original MCQ-style questions and their QA
 1128 counterparts.

1129

1130 The MCQ variants are used to measure baseline performance in the standard multiple-choice format,
 1131 which remains a common evaluation protocol for medical reasoning models. The QA variants are
 1132 used to evaluate models in an open-ended setting where answers are produced as ranked lists. In ad-
 1133 dition, we also employ QA benchmarks to evaluate free-form answers (without predefined options),
 1134 which serve as a more challenging baseline for assessing model generalization.

1135

1136 AlphaMed and QA-AlphaMed are used in RFT experiments, while SFT-* AlphaMed datasets are
 1137 used in SFT experiments. Mixed-AlphaMed is used for the training experiments described in Ap-
 1138 pendix K. Each record in Mixed-AlphaMed is prepended with an appropriate prior prompt, depend-
 1139 ing on the record type and the experiment. Additional details on how these datasets are used in train-

1134	Dataset	Split	Count	Task	License
Training datasets					
1137	AlphaMed (Liu et al., 2025a)	Train	19,778	MCQ	MIT
1138	QA-AlphaMed	Train	14,382	QA	Apache 2.0
1139	Mixed-AlphaMed	Train	34,160	Mixed	Apache 2.0
1140	SFT-MCQ-AlphaMed	Train	16,591	MCQ	Apache 2.0
1141	SFT-QA-AlphaMed	Train	9,416	QA	Apache 2.0
1142	SFT-List-AlphaMed	Train	9,705	QA	Apache 2.0
Evaluation benchmarks					
1144	MedQA ² (Jin et al., 2021)	Test	1,273	MCQ	MIT
1145	QA-MedQA	Test	1,233	QA	Apache 2.0
1146	MedMCQA ³ (Pal et al., 2022)	Test	6,150	MCQ	Apache 2.0
1147	QA-MedMCQA	Test	2,180	QA	Apache 2.0
1148	MedXpertQA ⁴ (Zuo et al., 2025)	Test-Text	2,450	MCQ	MIT
1149	QA-MedXpertQA	Test-Text	2,086	QA	Apache 2.0
1150	MMLU Pro ⁵ (Wang et al., 2024)	Test-Health	818	MCQ	MIT
1151	QA-MMLU Pro	Test-Health	736	QA	Apache 2.0

1152 Table 2: Overview of the training dataset and evaluation benchmarks. All datasets are publicly
 1153 available under licenses that permit their use for this type of research.

1154
 1155
 1156
 1157 ing and evaluation are provided in Section 2. The datasets are available at anonymous1entity/med-
 1158 datasets and anonymous1entity/med-sft-datasets.

1160 C MODEL OVERVIEW

1161
 1162 Table 3 summarizes the proprietary and open-weight models evaluated in this study. The proprietary
 1163 models (Gemini 2.5 family and GPT-4.1 Mini) do not disclose parameter counts, while the open-
 1164 weight models span several major families, including Qwen, Gemma, MedGemma, OpenThinker,
 1165 HuatuoGPT, m1, and AlphaMed. Importantly, all medical reasoning models sized at 7B parameters—
 1166 namely HuatuoGPT o1 7B, m1 7B 23K, and AlphaMed 7B Instruct RL—are derived from Qwen2.5
 1167 7B Instruct, with additional post-training targeted to medical domains. In contrast, MedGemma 4B
 1168 originates from Gemma 3 4B It.

1171 D PIPELINE FOR CONVERTING MCQs TO QA FORMAT

1172
 1173 This section describes the pipeline used to convert a dataset in MCQ format into QA format.
 1174 Inspired by Myrzakhan et al. (2024), we design a prompt to determine whether a given
 1175 question–choices–ground-truth triplet can be converted from MCQ to QA. First, we ask an LLM
 1176 to reason thoroughly before giving a verdict on whether the question can be converted. If the LLM
 1177 deems the conversion possible, it generates a QA-style question that yields the same ground-truth
 1178 answer.

1179
 1180 All of this happens in a single LLM call, since modern models show significant improvements in
 1181 performance, which reduces the need for a separate two-stage process. This simplification also
 1182 reduces the cost of conversion. We also ask the model to output confidence scores for further use in
 1183 the filtration process. We note that the number of resulting QA questions differs from the original
 1184 number of MCQs, as some questions may not be entirely suitable for conversion and are therefore
 1185 excluded. We use gpt-4.1-mini-2025-04-14 as the generation model with default sampling
 1186 parameters, except for the temperature, which we set to 0.1. The prompt used for conversion appears
 1187 in Figure 3. This pipeline applies to both the training dataset (AlphaMed) and the benchmarks used
 1188 in this study.

1188
 1189
 1190
 1191
 1192
 1193
 1194
 1195
 1196
 1197
 1198
 1199
 1200 Your task is to review a multiple-choice question, its answer choices, and the ground truth.
 1201 Determine if, after possible revision (including adding clarifying information), the question can be
 1202 answered correctly **without** the answer choices—as a standalone, open-ended question.
 1203 * For incomplete-sentence questions (e.g., "During swallowing, ..."), use your knowledge to
 1204 complete the sentence accurately.
 1205 * For identification questions (e.g., "Which of the following structures is part of the small intestine?")
 1206 , consider if the question can be revised so an informed respondent could answer it without choices.
 1207 * The revised question **MUST** be specific enough so that the answer can be determined without
 1208 ambiguity, and it **MUST BE** the ground truth.
 1209 If, even after revision, the question cannot be answered confidently without the choices, return **"**
 1210 **NO"**. If unsure, default to **"NO"**. Only return **"YES"** if you are confident the revised
 1211 question can be answered independently.
 1212 **Instructions:**
 1213 1. Reason through your decision inside '**<think>**' and '**</think>**' tags.
 1214 2. Output your verdict—**only** "YES" or "NO"—inside '**<verdict>**' and '**</verdict>**' tags.
 1215 Nothing else should appear within '**<verdict>**'.
 1216 3. If "YES", provide your revised version of the question inside '**<revised_question>**'
 1217 '**</revised_question>**'.
 1218 4. Finally, rate your confidence that this revised question can be answered in close-ended QA format
 1219 (1 = lowest, 5 = highest) inside '**<confidence>**'**</confidence>**'.
 1220 ---
 1221
 1222 **Question**
 1223 {question}
 1224
 1225 **Choices**
 1226 {choices}
 1227
 1228 **Ground truth**
 1229 {ground_truth}
 1230
 1231 Figure 3: Prompt used for converting MCQ questions into their equivalent QA format.
 1232
 1233
 1234
 1235
 1236
 1237
 1238
 1239
 1240
 1241

Model	Access	Size	Category
Gemini 2.5 Flash Lite ⁶ (Comanici et al., 2025)	P	N/A	GRM
Gemini 2.5 Flash ⁷ (Comanici et al., 2025)	P	N/A	GRM
Gemini 2.5 Pro ⁸ (Comanici et al., 2025)	P	N/A	GRM
GPT-4.1 Mini ⁹ (OpenAI, 2025)	P	N/A	GLM
Qwen2.5 3B Instruct ¹⁰ (Qwen et al., 2025)	OW	3B	GLM
Qwen2.5 7B Instruct ¹¹ (Qwen et al., 2025)	OW	7B	GLM
Qwen2.5 14B Instruct ¹² (Qwen et al., 2025)	OW	14B	GLM
Qwen3 4B Instruct 2507 ¹³ (Yang et al., 2025)	OW	4B	GLM
Gemma 3 4B IT ¹⁴ (Team et al., 2025)	OW	4B	GLM
MedGemma 4B IT ¹⁵ (Sellergren et al., 2025)	OW	4B	MLM
MedGemma 27B IT ¹⁶ (Sellergren et al., 2025)	OW	27B	MLM
OpenThinker 3 7B ¹⁷ (Guha et al., 2025)	OW	7B	GRM
HuatuoGPT 01 7B ¹⁸ (Chen et al., 2024)	OW	7B	MRM
m1 7B 23K ¹⁹ (Huang et al., 2025a)	OW	7B	MRM
AlphaMed 7B Instruct RL ²⁰ (Liu et al., 2025a)	OW	7B	MRM

Table 3: Overview of **proprietary (P)** and **open-weight (OW)** models used in this study, categorized as **general reasoning models (GRM)**, **medical reasoning models (MRM)**, **general-purpose LLMs (GLM)**, or **medical-specialized LLMs (MLM)**.

E EVALUATION METRICS

E.1 ADDITIONAL DETAILS ON METRICS

For the MCQ (Acc_{MCQ}) setting, accuracy is computed by exact match between the predicted choice and the ground truth. For the QA (Acc_{QA}) setting, we use normalized (lowercasing) exact match between the extracted answer and the ground truth to obtain accuracy. For the ranked-list setting, we report both accuracy (Acc_{List})—whether the ground-truth answer appears anywhere in the list—and MRR (MRR_{List}), which additionally accounts for the position of the correct answer, assigning higher scores when it appears earlier in the list.

Since models may produce correct answers that do not exactly match the ground-truth string, we also utilize LLM-based evaluation variants for non-MCQ formats: LLM-Acc for QA (Acc_{QA}^{LLM}) and ranked lists (Acc_{List}^{LLM}) and LLM-MRR (MRR_{List}^{LLM}) for ranked lists. These provide more flexible judgments of correctness and complement the exact-match metrics. Reporting both exact-match and LLM-based metrics provides a more nuanced perspective on training effects and generalization.

E.2 LLM-BASED METRICS

There are three LLM-based metrics utilized in this study: LLM-Acc for QA answers, LLM-Acc for list answers, and LLM-MRR for list answers. These LLM-based metrics can capture semantically correct answers that differ in surface form, our hypothesis is that effective training should already improve performance under strict exact-match evaluation, even without relying on the more forgiving LLM-based measures. All LLM-based metrics use `gpt-4.1-mini-2025-04-14` as the judge. The prompt used for LLM-Acc on QA answers is shown in Figure 4, while the prompt for LLM-MRR on list answers is shown in Figure 5.

These prompts incorporate validation rules to guard against attempts to game the judge by producing nonsensical answers. We also instruct the model to respond in a predefined format to facilitate answer extraction. All evaluations are performed with the default sampling temperature, except when explicitly set to 0.0.

We note that LLM-Acc is derived from the results of the LLM-MRR judge. In particular, an LLM receives an LLM-Acc score of 1.0 for a given question if the correct answer appears in the output,

1296
1297
1298
1299
1300
1301
1302

1303 You are evaluating whether predicted answers match the ground truth answer semantically, even if
1304 they are not exactly the same text.

1305

1306 Ground Truth Answer: "{ground_truth}"

1307

1308 Predicted Answers: {predicted_text}

1309

1310 **IMPORTANT VALIDATION RULES:**

1311

1. Each predicted answer must contain EXACTLY ONE focused response
2. If any answer contains multiple distinct answers, options, or attempts to cover multiple possibilities (e.g., using "or", "and", commas to separate different answers, bullet points, or lists), that answer is INVALID
3. Too long answers will be considered invalid gaming attempts
4. Only evaluate answers that contain single, focused responses
5. Answers containing repeated words, phrases, or synonyms (e.g., "pneumonia pneumonia", "MI heart attack myocardial infarction", "diabetes DM diabetes mellitus") are INVALID
6. More than 2 occurrences of the same root word or concept in a single answer is INVALID
7. Overly broad terms that could match multiple conditions (e.g., "infection", "cardiac event", "abnormal values", "inflammatory condition") are INVALID unless they precisely match the ground truth
8. Use of special characters, symbols, or separators to bypass comma rules (2.) (e.g., "|", "/", ":";, ";", unicode spaces) is INVALID
9. Incomplete answers containing only part of medical terms or abbreviations without full context are INVALID
10. Answers containing non-medical terms, gibberish, or obvious filler content are INVALID
11. If multiple answers contain very similar but slightly different versions of the same answer, only the first valid answer should be considered
12. Medically impossible or contradictory answers indicate INVALID
13. The same medical concept appearing in multiple answers with slight variations is INVALID

1312

1313 Task: Determine if ANY of the predicted answers is semantically equivalent to the ground truth
1314 answer. Consider:

1315

- Different phrasings that mean the same thing
- Medical terminology and abbreviations
- Minor spelling variations
- Synonyms and equivalent expressions
- Clinical accuracy and meaning
- Answers must be clinically specific enough to distinguish from other conditions
- Abbreviations must be standard medical abbreviations, not made-up shortcuts
- Technical terms must be used correctly in proper medical context
- Synonyms must be legitimate medical synonyms, not loose associations

1316

1317 **RESPONSE FORMAT:**

1318

1319 Provide your reasoning and analysis before giving the final verdict. Keep it short and concise.

1320

1321 Verdict: [Provide ONLY "MATCH" if any predicted answer is semantically equivalent to the ground

1322

1323 truth, or "NO_MATCH" if none are equivalent or if gaming behavior is detected]

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

Figure 4: Prompt used for evaluating LLM-Acc on QA answers.

1350
 1351
 1352
 1353
 1354
 1355
 1356 You are evaluating which predicted answer (if any) matches the ground truth answer semantically.
 1357
 1358 Ground Truth Answer: "{ground_truth}"
 1359 Predicted Answers (in rank order):
 1360 {predicted_list}
 1361
IMPORTANT VALIDATION RULES:
 1362 1. Each rank position must contain EXACTLY ONE focused answer
 1363 2. If any rank contains multiple distinct answers, options, or attempts to cover multiple possibilities (e.g., using "or", "and", commas to separate different answers, bullet points, or lists), that rank is
 1364 INVALID
 1365 3. Too long answers will be considered invalid gaming attempts
 1366 4. Only evaluate ranks that contain single, focused responses
 1367 5. Ranks containing repeated words, phrases, or synonyms (e.g., "pneumonia pneumonia", "MI heart
 1368 attack myocardial infarction", "diabetes DM diabetes mellitus") are INVALID
 1369 6. More than 2 occurrences of the same root word or concept in a single rank is INVALID
 1370 7. Overly broad terms that could match multiple conditions (e.g., "infection", "cardiac event", "abnormal values", "inflammatory condition") are INVALID unless they precisely match the ground
 1371 truth
 1372 8. Use of special characters, symbols, or separators to bypass comma rules (2.) (e.g., "|", "/", ":";, ";;",
 1373 unicode spaces) is INVALID
 1374 9. Incomplete answers containing only part of medical terms or abbreviations without full context are
 1375 INVALID
 1376 10. Answers containing non-medical terms, gibberish, or obvious filler content are INVALID
 1377 11. If lower ranks (2, 3, 4, etc.) contain obviously incorrect answers that appear designed to make
 1378 rank 1 look better by comparison, treat as INVALID
 1379 12. If multiple ranks contain very similar but slightly different versions of the same answer, return
 1380 the first valid rank
 1381 13. Medically impossible or contradictory answers across ranks indicate INVALID
 1382 14. The same medical concept appearing in multiple ranks with slight variations is INVALID
 1383
 Task: Determine which predicted answer number (1, 2, 3, etc.) is semantically equivalent to the
 1384 ground truth answer. Consider:
 1385 – Different phrasings that mean the same thing
 1386 – Medical terminology and abbreviations
 1387 – Minor spelling variations
 1388 – Synonyms and equivalent expressions
 1389 – Clinical accuracy and meaning
 1390 – Answers must be clinically specific enough to distinguish from other conditions
 1391 – Abbreviations must be standard medical abbreviations, not made-up shortcuts
 1392 – Technical terms must be used correctly in proper medical context
 1393 – Synonyms must be legitimate medical synonyms, not loose associations
 1394
RESPONSE FORMAT:
 1395 Provide your reasoning and analysis before giving the final verdict. Keep it short and concise.
 1396 Verdict: [Provide ONLY the number of the first valid matching answer (e.g., "1", "2", "3", etc.), or "NO_MATCH" if none are equivalent or if gaming behavior is detected]
 1397
 1398 Figure 5: Prompt used for evaluating LLM-MRR on list answers.
 1399
 1400
 1401
 1402
 1403

1404
 1405 You are evaluating which predicted answer (if any) matches the ground truth answer semantically.
 1406
 1407 Ground Truth Answer: ”{ground_truth}”
 1408
 1409 Predicted Answers (in rank order):
 1410 {predicted_list}
 1411
 1412 Task: Determine which predicted answer number (1, 2, 3, etc.) is semantically equivalent to the
 1413 ground truth answer. Consider:
 1414 – Different phrasings that mean the same thing
 1415 – Medical terminology and abbreviations
 1416 – Minor spelling variations
 1417 – Synonyms and equivalent expressions
 1418 – Clinical accuracy and meaning
 1419
 1420 Respond with ONLY the number of the first matching answer (e.g., ”1”, ”2”, ”3”, etc.), or ”
 1421 NO_MATCH” if none are equivalent.

Figure 6: A simpler judge prompt used for an ablation study in Section 4.3.1.

1422
 1423
 1424 You are a helpful and harmless expert clinical assistant. The assistant provides the user with the
 1425 accurate answer. When you finally reach a conclusion, clearly state the final answer in \boxed{ }.
 1426 You always begins your answer with the choice, e.g., A., B., C., D., E. in the \boxed{ } . Now the user
 1427 asks you to solve a problem.
 1428
 1429 {query}
 1430 {choices}

Figure 7: Zero-shot prompt template for multiple-choice questions (MCQ).

1431
 1432 regardless of its rank. This contrasts with LLM-MRR, where the rank assigned by the judge is also
 1433 taken into account when computing the reward.

1434 The prompt for LLM-MRR is also used as a judge prompt during RFT in Section 4.2 as well. We
 1435 also have a simpler version of this judge prompt, which is Figure 6, used for an ablation study in
 1436 Section 4.3.1.

1437 F PROMPT TEMPLATES

1438 We design a total of six prompt templates: MCQ, MCQ-CoT, QA, QA-CoT, List, and List-CoT.
 1439 Our templates are adapted from the prior-prompt approach introduced by Xie et al. (2025), with
 1440 modifications to better suit the medical domain. Specifically, we adjust the role prompt and, in the
 1441 MCQ-CoT variant, explicitly require the model to begin its final answer with the selected choice
 1442 marker (e.g., A, B, C). The other answer formats reuse the same general template with instructions
 1443 tailored to QA or list-style outputs. For the list format, we additionally provide a one-shot example
 1444 to illustrate the expected output structure.

1445 For zero-shot variants, we remove the reasoning instruction and omit the `<think>` and `</think>`
 1446 tags, leaving only the answer-format instruction and general components. All prompting ex-
 1447 periments are run with consistent decoding parameters across models to ensure comparability:
 1448 `temperature=0.0`, `top_p=1.0`, and `top_k=-1`. We set `max_tokens=8192` for most mod-
 1449 els, but increase this to 16384 for Gemini 2.5 models to accommodate their typically more verbose
 1450 reasoning chains. The complete set of prompt templates used in our experiments is shown in Fig-
 1451 ures 7 to 12.

1458

1459

1460 ou are a helpful and harmless expert clinical assistant. The assistant first thinks about the reasoning
 1461 process in the mind and then provides the user with the accurate answer. The reasoning process is
 1462 enclosed within `<think></think>` tags followed by an answer, i.e., `<think>` reasoning process here
 1463 `</think>` answer here. After thinking, when you finally reach a conclusion, clearly state the final
 1464 answer in `\boxed{}`. You always begins your answer with the choice, e.g., A., B., C., D., E. in the `\boxed{}`. Now the user asks you to solve a problem.

1465

1466

{query}
 {choices}

1468

1469

1470 Figure 8: CoT prompt template for multiple-choice questions (MCQ-CoT).

1471

1472

1473

1474 You are a helpful and harmless expert clinical assistant. The assistant provides the user with the
 1475 accurate answer. Now the user asks you to solve a problem. When you finally reach a conclusion,
 1476 clearly state the final answer in `\boxed{}`.

1477

1478

{query}

1479

1480

1481

1482

1483

1484

1485 You are a helpful and harmless expert clinical assistant. The assistant first thinks about the reasoning
 1486 process in the mind and then provides the user with the accurate answer. The reasoning process is
 1487 enclosed within `<think></think>` tags followed by an answer, i.e., `<think>` reasoning process here
 1488 `</think>` answer here. After thinking, when you finally reach a conclusion, clearly state the final
 1489 answer in `\boxed{}`. Now the user asks you to solve a problem.

1490

1491

1492

1493

Figure 9: Zero-shot prompt template for open-ended QA.

1494

1495

1496

1497 You are a helpful and harmless expert clinical assistant. The assistant provides the user with an
 1498 accurate answer. When you finally reach a conclusion, clearly list all possible answers in order from
 1499 most likely to least likely. Start with "# Final Answer" followed by numbered lines using the format 'n. answer' for each answer. Each item MUST contain only the answer without any explanation or
 1500 reasoning.

1501

1502

Example:

Final Answer
 1. xxx
 2. xxx

1505

1506

Now the user asks you to solve a problem.

1507

1508

{query}

1509

1510

1511

Figure 11: Zero-shot prompt template for list-style answers.

1512
1513
1514
1515
1516
1517
1518
1519

You are a helpful and harmless expert clinical assistant. The assistant first thinks about the reasoning process and then provides the user with an accurate answer. The reasoning process is enclosed within `<think></think>` tags followed by an answer, i.e., `<think>reasoning process here</think>` answer here. After thinking, when you finally reach a conclusion, clearly list all possible answers in order from most likely to least likely. Start with "# Final Answer" followed by numbered lines using the format 'n. answer' for each answer. Each item MUST contain only the answer without any explanation or reasoning.

1520
1521
1522
1523
1524
1525

Example:
`<think>...</think>`

Final Answer
 1. xxx
 2. xxx

1526
1527
1528
1529

Now the user asks you to solve a problem.

{query}

1530
1531
1532

Figure 12: CoT prompt template for list-style answers (List-CoT).

1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547

You are a medical validation expert. Your task is to validate whether a medical response contains the correct answer.

Given:

- Correct Answer: `<CORRECT_ANSWER>`
- Generated Response: `<RESPONSE>`

Please determine if the generated response contains or aligns with the correct answer. Consider:

1. For MCQ questions with option letters (A, B, C, D, E): Check if the response contains the correct option letter, and optional answer
2. For other questions: Check if the response mentions the correct answer explicitly or implicitly

Respond with only "VALID" if the response contains the correct answer, or "INVALID" if it does not.

1548
1549
1550

Figure 13: Validation prompt used for multiple-choice (MCQ) and multiple-answer (MQA) formats.

1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

G SFT TRAINING DATASET PREPARATION

We construct the SFT training dataset through knowledge distillation from Qwen3-30B-A3B-Thinking-2507-FP8²¹ (Yang et al., 2025). Specifically, we provide questions from AlphaMed and AlphaMedQA, coupled with the CoT variants of each prompt template corresponding to the answer format under consideration.

To ensure correctness, we apply rejection sampling using an LLM judge, gpt-4o-mini-2024-07-18. For MCQ and MQA responses, we use the validation prompt shown in Figure 13, while list answers are validated with the prompt in Figure 14. We use a sampling temperature of 0.7, a maximum token length of 8192, and allow up to 20 retries for incorrect responses. Responses that remain incorrect after rejection sampling are discarded. The filtered records are retained and used to train the distilled SFT models. Additional details on training are provided in Appendix H.1.

²¹<https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507-FP8>

1566
 1567
 1568 You are a medical validation expert. Your task is to validate whether a medical response with a list
 1569 format contains the correct answer.
 1570 Given:
 1571 – Correct Answer: <CORRECT_ANSWER>
 1572 – Generated Response: <RESPONSE>
 1573
 1574 The generated response should contain a numbered list of possible answers. Please determine if the
 1575 correct answer appears anywhere in this list. Consider:
 1576 1. The correct answer may appear as an exact match in one of the list items
 1577 2. The correct answer may appear with slight variations or paraphrasing
 1578 3. Look for the answer in the "# Final Answer" section with numbered items
 1579
 1580 Respond with only "VALID" if the correct answer appears in the generated list, or "INVALID" if it
 1581 does not.
 1582

Figure 14: Validation prompt used for list-style answers.

H TRAINING HYPERPARAMETERS

In this section, we describe the training hyperparameters used in our experiments. All training experiments used about 815 GPU hours on a 4xH100 node.

H.1 SFT

We use LLaMA-Factory²² (Zheng et al., 2024) v0.9.3, which is released under the Apache 2.0 license. For SFT, we perform full fine-tuning with DeepSpeed ZeRO Stage 3 (Rajbhandari et al., 2020). Training is conducted with a per-device batch size of 2 and a gradient accumulation step of 4, resulting in an effective batch size of 8. We use a learning rate of 1×10^{-5} for 2 epochs with a cosine learning rate scheduler. Training is performed with `b16` precision, and FlashAttention-2 is enabled to improve efficiency. We set the warmup ratio to 0.05. The training datasets are the SFT-*‐AlphaMed variants listed in Table 2. Each variant is used to train one model, resulting in three models in total, corresponding to the three answer formats investigated in this study.

H.2 RFT

We use verl²³ (Sheng et al., 2025) v0.5.0, released under the Apache 2.0 license. Training is performed with full fine-tuning (no offloading) using FSDP2 (Zhao et al., 2023) as the backend and group relative policy optimization (GRPO) (Shao et al., 2024) without a KL regularization term (Liu et al., 2025b; Xie et al., 2025).

We train with a batch size of 256, divided into 64 mini-batches. The maximum prompt length is 2048 tokens, and the maximum response length is 4096 tokens, constrained by available compute. The learning rate is set to 1×10^{-6} , with padding removed and gradient checkpointing enabled. Torch compile is also enabled for efficiency.

For rollouts, we use vLLM²⁴ (Kwon et al., 2023). Log-probability computation is performed with a micro-batch size of 8 per GPU. For each prompt, we sample 8 responses with the default verl parameters. Dynamic batching is enabled for greater efficiency, targeting a maximum of 24,576 tokens per GPU for the actor, reference, and rollout models. We train for 2 epochs with no warmup. The initial model, training set, and reward function are selected according to the configuration of each experiment. The reward function code is available in the repository.

²²<https://github.com/hiyouga/LLaMA-Factory>

²³<https://github.com/volgengine/verl>

²⁴<https://github.com/vllm-project/vllm>

1620 Table 4: Average response length (mean \pm standard deviation) for MCQ, QA, and list-based answer
 1621 formats across benchmarks. Note that **LL** denotes a list length, i.e., the mean number of items across
 1622 lists, including empty lists (items = 0.)

	MCQ	QA	List	CP	LL	VLL
<i>Proprietary Models</i>						
Gemini 2.5 Flash Lite	1585 \pm 3720	511 \pm 997	93 \pm 735	1.39	2.77	2.86
+CoT	4360.07 \pm 5704	2850 \pm 4535	7398 \pm 8135	1.17	0.87	1.92
Gemini 2.5 Flash	473 \pm 251	273 \pm 220	32 \pm 33	1.41	2.98	3.00
+CoT	1759 \pm 1367	1576 \pm 1036	1021 \pm 857	1.30	1.16	2.66
Gemini 2.5 Pro	425 \pm 188	459 \pm 289	38 \pm 55	1.40	3.38	3.41
+CoT	1326 \pm 326	1527 \pm 568	1019 \pm 282	1.41	3.29	3.46
GPT-4.1 Mini	353 \pm 156	285 \pm 158	137 \pm 86	1.36	3.20	3.26
+CoT	357 \pm 122	292 \pm 110	252 \pm 88	1.36	3.71	3.72
<i>Open-weight Models</i>						
Qwen2.5 3B Instruct	271 \pm 221	214 \pm 254	36 \pm 32	1.69	3.09	3.09
+CoT	361 \pm 305	315 \pm 385	167 \pm 234	1.27	1.78	1.80
Qwen2.5 7B Instruct (our initial model)	72 \pm 68	146 \pm 138	27 \pm 18	1.45	2.39	2.39
+CoT	2393 \pm 3355	196 \pm 125	4434 \pm 3991	1.91	184.19	185.72
Qwen2.5 14B Instruct	129 \pm 72	140 \pm 124	35 \pm 21	1.48	3.04	3.05
+CoT	217 \pm 123	213 \pm 237	163 \pm 196	1.30	2.36	2.36
Qwen3.4B Instruct 2507	757 \pm 867	450 \pm 660	69 \pm 214	1.70	3.96	3.96
+CoT	894 \pm 929	538 \pm 646	458 \pm 808	1.50	3.32	3.56
Gemma 3.4B IT	400 \pm 112	432 \pm 200	61 \pm 135	1.83	4.59	4.68
+CoT	382 \pm 669	304 \pm 163	251 \pm 527	1.67	3.43	4.13
MedGemma 4B IT	297 \pm 809	152 \pm 447	951 \pm 2554	2.15	95.01	95.06
+CoT	7958 \pm 1141	493 \pm 1445	8180 \pm 309	3.14	419.18	482.76
MedGemma 27B IT	1081 \pm 847	653 \pm 636	52 \pm 214	1.46	3.20	3.26
+CoT	1424 \pm 878	966 \pm 949	1016 \pm 1050	1.43	3.81	3.88
OpenThinker3 7B	4789 \pm 2695	4348 \pm 2844	4744 \pm 2982	1.43	2.06	3.06
+CoT	7340 \pm 2139	4450 \pm 2862	7630 \pm 1729	1.29	0.23	4.76
HuatuoGPT o1 7B	472 \pm 169	492 \pm 260	47 \pm 224	1.70	2.80	4.39
+CoT	488 \pm 210	501 \pm 202	375 \pm 272	1.46	0.03	2.45
m1 7B 23K	1578 \pm 2203	1542 \pm 1967	1657 \pm 2930	2.01	13.39	13.39
+CoT	8185 \pm 202	2328 \pm 2423	8149 \pm 542	1.64	14.62	19.03
AlphaMed 7B Instruct RL	311 \pm 187	216 \pm 275	19 \pm 63	1.83	1.93	2.59
+CoT	416 \pm 480	266 \pm 607	1052 \pm 2503	1.89	47.86	55.74
<i>Our Knowledge-Distilled Medical Reasoning Models (based on Qwen2.5 7B Instruct)</i>						
SFT-MCQ	2749 \pm 1626	1419 \pm 1204	2438 \pm 3310	2.83	141.11	141.72
+CoT	2643 \pm 1577	1390 \pm 1203	1671 \pm 1277	1.09	1.44	1.46
SFT-QA	2365 \pm 1646	1235 \pm 1045	1758 \pm 2355	1.81	13.50	13.85
+CoT	2425 \pm 1605	1388 \pm 1136	15312 \pm 1172	1.15	1.04	1.78
SFT-List	2856 \pm 1614	1287 \pm 1184	1281 \pm 992	1.41	2.50	2.52
+CoT	2776 \pm 1591	1333 \pm 1153	1425 \pm 1112	1.42	2.55	2.57
<i>Our RFT Medical Reasoning Models (based on Qwen2.5 7B Instruct)</i>						
RFT-MCQ	204 \pm 106	167 \pm 65	33 \pm 123	1.45	2.29	2.29
RFT-QA	296 \pm 302	291 \pm 247	46 \pm 91	1.67	0.29	3.01
RFT-List-Acc	208 \pm 231	195 \pm 433	165 \pm 177	2.07	5.94	5.94
RFT-List-MRR	174 \pm 110	157 \pm 207	319 \pm 1154	2.11	16.97	16.97
RFT-List-Judge-MRR	168 \pm 376	193 \pm 622	133 \pm 214	1.64	4.45	4.46

I ADDITIONAL DISCUSSIONS

In this section, we provide an additional discussion on results presented in the main paper.

I.1 PROMPTING RESULTS

MCQ vs. QA For HuatuoGPT-o1 and AlphaMed, the drop is expected, as both were trained specifically on MCQ. Interestingly, this is not the case for m1, which was also trained on MCQ data for medical benchmarks. Although HuatuoGPT-o1 and m1 used comparable training data sizes (\approx 20K samples), their training paradigms differ. Both HuatuoGPT-o1 and m1 were trained with SFT, whereas AlphaMed was trained with RFT. The key distinction between HuatuoGPT-o1 and m1 is that the former relied on synthetic reasoning trajectories, while m1 utilized distilled trajectories from a large reasoning model.

Non-MCQ formats are unfamiliar to the models We conduct a Wilcoxon signed-rank test²⁵ comparing MCQ with other formats across benchmarks, prompting strategies, and models. The results show that changing the answer format from *MCQ* to *QA* ($p = 4.26 \times 10^{-4}$, $r = 0.317$) significantly alters model performance, with a small-to-moderate effect size. The effect is even

²⁵We found evidence against normality for both pairs using the Shapiro–Wilk test ($p = 0.000614$ for *MCQ* vs. *QA* and $p = 0.0016$ for *MCQ* vs. *List*). Consequently, we employed the Wilcoxon signed-rank test.

1674 stronger for *MCQ* to *ranked-list* ($p = 9.38 \times 10^{-16}$, $r = 0.667$), indicating a large effect. We
 1675 attribute this to the fact that the majority of current medical benchmarks are available in *MCQ*
 1676 format (e.g., (Jin et al., 2021; Pal et al., 2022; Zuo et al., 2025; Wang et al., 2024)), and models
 1677 trained to excel in this setting may have learned to associate the answer format with knowledge
 1678 (Li et al., 2024; Singh et al., 2025). Consequently, changing the answer format leads to substantial
 1679 performance differences, particularly for the *ranked-list* format, which is less common and thus less
 1680 familiar to models.

1681
 1682 **Do longer responses lead to better performance?** We observe from Table 4 that the majority of
 1683 reasoning models produce longer responses than standard LLMs. While this trend holds for most
 1684 reasoning models (for example, OpenThinker3 and m1 generate long responses of around 4K and
 1685 1K tokens, respectively, regardless of answer format), AlphaMed is an exception. AlphaMed is the
 1686 only open-weight medical reasoning model trained with RFT rather than SFT. We further discuss
 1687 the impact of RFT on response length in Section 4.2.

1688 Statistical testing using Pearson’s correlation between model score and mean response length across
 1689 all prompting variants, benchmarks, metrics, answer formats, and models included in this experi-
 1690 ment revealed a small but statistically significant negative correlation ($r = -0.144$, $p = 7.4 \times 10^{-6}$,
 1691 $|r| = 0.144$). These results indicate that performance is negatively, but only weakly, associated
 1692 with response length. In other words, producing longer responses weakly and negatively affects
 1693 performance.

1694 Therefore, the premise that reasoning models always produce longer answers (Guo et al., 2025;
 1695 Muennighoff et al., 2025) and that longer responses signal greater performance (Guo et al., 2025)
 1696 does not hold in our setting. In fact, prior work on efficient reasoning (Sui et al., 2025) suggests
 1697 that various training techniques can encourage concise reasoning chains while maintaining high
 1698 performance. This further supports the conclusion that response length is a poor indicator of final
 1699 model performance.

1700 I.2 FINE-TUNING RESULTS

1702 I.3 SFT RESULTS

1704 As shown in Table 4, models trained with SFT in any format consistently produce longer responses
 1705 (around 1K–3K tokens). These findings align with what was observed with m1, another knowledge-
 1706 distilled medical reasoning model from prior work, discussed in Section 3.

1707 I.4 RFT RESULTS

1709 **RFT does not always incentivize long responses** Models trained with RFT are surprisingly con-
 1710 cise, often producing shorter answers than the initial model prompted with CoT, similar to Al-
 1711 phaMed and HuatouGPT o1. We conjecture that RFT primarily incentivizes models to make the
 1712 most effective use of their intermediate generated tokens to maximize performance, rather than en-
 1713 couraging longer responses. However, longer responses may still correlate with higher accuracy in
 1714 certain scenarios.

1716 I.5 ABLATION STUDIES FOR RFT

1718 I.5.1 RFT FACTORS

1719 Tables 5 and 6 present the performance and list metrics for the experiments in Section 4.3.1. In the
 1720 no-prior-prompt setting, models tend to generate longer responses (e.g., RFT-MCQ averages 204 →
 1721 474 tokens), whereas responses are slightly shorter under the no-CoT-prompt setting (e.g., 204 →
 1722 129 tokens). By contrast, both judge models yield similar list behaviors and training dynamics, such
 1723 as average response lengths around 168–239 tokens for MCQ and 133–141 tokens for list outputs.

1724 Effects of Prior Prompts To account for changes in prior prompts, we adjust our setup when remov-
 1725 ing prior prompts or parts of them. Specifically, in the no-prompt and no-CoT settings, we exclude
 1726 the format component from the reward function, since the absence of explicit thinking tags would
 1727 otherwise drive it toward zero.

1728
1729
1730
1731
1732
1733 Table 5: Performance results of the ablation study on factors affecting RFT. The focus is on the re-
1734 ward component in the reward function, extended training duration, and the effects of prior prompts
1735 across models. **Rw.Fn.** denotes Reward Function.

	Prior Prompt	Rw.Fn.	MCQ		QA		List		
			Acc _{MCQ}	Acc _{QA}	Acc _{QA} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}
RFT-MCQ	MCQ-CoT	Acc _{MCQ}	39.34	9.96	46.33	9.89	40.06	8.16	33.00
No format reward	MCQ-CoT	Acc _{MCQ-NF}	39.56	9.48	46.24	10.01	37.91	9.19	33.47
4 Epochs	MCQ-CoT	Acc _{MCQ}	39.97	9.78	45.07	9.91	35.46	9.35	32.20
No prompt	X	Acc _{MCQ-NF}	38.95	9.91	46.12	10.32	41.94	9.25	36.49
MCQ prompt	MCQ	Acc _{MCQ-NF}	39.80	9.21	44.59	10.04	38.78	8.91	32.73
RFT-QA	QA-CoT	Acc _{QA}	36.80	1.04	25.22	0.62	3.59	0.48	2.82
No prompt	X	Acc _{QA-NF}	27.76	4.08	29.17	9.75	41.92	8.32	34.75
QA prompt	QA	Acc _{QA-NF}	28.23	0.95	24.78	11.42	49.93	9.06	38.35
RFT-List-Acc	List-CoT	Acc _{List}	22.40	4.28	19.01	22.11	56.61	16.17	40.26
List prompt	List	Acc _{List-NF}	12.97	9.43	44.46	24.16	67.08	13.83	37.13
RFT-List-MRR	List-CoT	MRR _{List}	18.23	4.43	21.90	20.96	61.60	15.83	44.89
List prompt	List	MRR _{List-NF}	10.41	10.37	46.21	22.06	63.00	15.28	41.24
RFT-List-Judge-MRR	List-CoT	MRR _{List} ^{LLM}	20.49	6.49	30.36	14.86	60.90	12.16	48.68
Gemini judge	List-CoT	MRR _{List} ^{LLM-Gemini}	33.11	9.40	43.16	13.07	59.34	10.95	48.00
Simple judge prompt	List-CoT	MRR _{List} ^{LLM-Simple}	19.02	4.37	27.27	4.75	31.84	3.86	26.19

1751
1752
1753
1754
1755
1756
1757
1758
1759
1760 Table 6: Average response length (mean \pm standard deviation) for MCQ, QA, and list-based answer
1761 formats across benchmarks and metrics related to the ranked list answer format from the generated
1762 evaluation responses for the RFT factors ablation study.

	Prior Prompt	Rw.Fn.	MCQ		QA		List	CP	LL	VLL
			204 \pm 106	167 \pm 65	33 \pm 123	1.45	2.29	2.29	1.31	1.96
RFT-MCQ	MCQ-CoT	Acc _{MCQ}	204 \pm 106	167 \pm 65	33 \pm 123	1.45	2.29	2.29	1.31	1.96
No format reward	MCQ-CoT	Acc _{MCQ-NF}	178 \pm 138	159 \pm 57	129 \pm 114	1.31	1.96	1.97	1.24	1.74
4 Epochs	MCQ-CoT	Acc _{MCQ}	336 \pm 107	211 \pm 138	178 \pm 141	1.34	2.23	2.23	1.34	2.23
No prompt	X	Acc _{MCQ-NF}	474 \pm 547	322 \pm 388	147 \pm 196	1.34	2.23	2.23	1.34	2.23
MCQ prompt	MCQ	Acc _{MCQ-NF}	129 \pm 112	121 \pm 79	24 \pm 16	1.41	2.14	2.14	1.41	2.14
RFT-QA	QA-CoT	Acc _{QA}	296 \pm 302	291 \pm 247	46 \pm 91	1.67	0.29	3.01	1.50	2.80
No prompt	X	Acc _{QA-NF}	259 \pm 378	414 \pm 1015	156 \pm 288	1.68	3.48	3.48	1.68	3.48
QA prompt	QA	Acc _{QA-NF}	130 \pm 183	158 \pm 207	43 \pm 22	1.68	3.48	3.48	1.68	3.48
RFT-List-Acc	List-CoT	Acc _{List}	208 \pm 231	195 \pm 433	165 \pm 177	2.07	5.94	5.94	5.57	615.17
List prompt	List	Acc _{List-NF}	85 \pm 114	144 \pm 194	7864 \pm 1551	5.57	615.17	615.17		
RFT-List-MRR	List-CoT	MRR _{List}	174 \pm 110	157 \pm 207	319 \pm 1154	2.11	16.97	16.97	2.59	9.51
List prompt	List	MRR _{List-NF}	28 \pm 75	100 \pm 150	79 \pm 235	2.59	9.51	9.51		
RFT-List-Judge-MRR	List-CoT	MRR _{List} ^{LLM}	168 \pm 376	193 \pm 622	133 \pm 214	1.64	4.45	4.46	1.58	4.25
Gemini judge	List-CoT	MRR _{List} ^{LLM-Gemini}	239 \pm 131	200 \pm 142	141 \pm 103	1.58	4.25	4.25	1.58	4.25
Simple judge prompt	List-CoT	MRR _{List} ^{LLM-Simple}	192 \pm 194	163 \pm 286	140 \pm 43	1.51	3.53	3.53	1.51	3.53

1782
1783
1784 Table 7: Performance results of the ablation study on different initial models.
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835

	MCQ		QA		List		
	Acc _{MCQ}	Acc _{QA}	Acc _{QA} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}
RFT-MCQ							
Qwen2.5 7B Instruct	39.34	9.96	46.33	9.89	40.06	8.16	33.00
Qwen2.5 3B Instruct	31.28	6.82	36.45	7.74	32.20	6.96	28.01
Qwen3 4B Instruct	45.22	11.00	46.05	13.66	54.82	12.01	46.18
OpenThinker3 7B	33.74	5.19	40.85	6.88	33.28	6.27	29.11
m1 7B 23K	44.98	8.49	41.67	12.40	46.24	11.20	40.49
AlphaMed 7B	42.03	3.50	18.64	1.52	3.57	1.51	3.44
RFT-QA							
Qwen2.5 7B Instruct	36.80	1.04	25.22	0.62	3.59	0.48	2.82
Qwen2.5 3B Instruct	32.64	0.84	27.60	9.84	44.78	7.34	33.94
Qwen3 4B Instruct	45.16	8.56	44.97	12.34	55.46	10.53	46.57
OpenThinker3 7B	34.01	4.67	39.78	4.31	39.02	3.90	33.49
m1 7B 23K	43.06	6.88	45.54	9.53	45.35	8.52	40.04
AlphaMed 7B	41.74	3.94	38.35	10.14	37.89	9.63	34.95
RFT-List-Acc							
Qwen2.5 7B Instruct	22.40	4.28	19.01	22.11	56.61	16.17	40.26
Qwen2.5 3B Instruct	32.16	7.25	39.29	20.66	59.82	12.08	34.68
Qwen3 4B Instruct	43.72	11.69	48.45	27.74	71.60	15.94	40.60
OpenThinker3 7B	34.23	5.75	41.42	20.05	56.98	11.03	31.93
m1 7B 23K	44.04	8.96	43.92	26.48	66.34	15.52	37.11
AlphaMed 7B	38.35	5.82	25.38	17.86	57.29	14.32	44.16
RFT-List-MRR							
Qwen2.5 7B Instruct	18.23	4.43	21.90	20.96	61.60	15.83	44.89
Qwen2.5 3B Instruct	32.01	7.93	40.20	20.52	58.42	12.53	34.91
Qwen3 4B Instruct	44.32	11.93	48.54	17.60	58.60	12.32	40.27
OpenThinker3 7B	34.60	5.57	40.77	18.23	59.44	11.25	35.03
m1 7B 23K	43.54	9.85	45.36	23.57	67.29	15.61	42.10
AlphaMed 7B	36.44	7.15	30.38	18.47	55.22	15.05	43.91

No prior prompt The no-prior-prompt setting is only applicable to MCQ and QA, since a ranked-list format require a one-shot example. For RFT-MCQ setup, Acc_{MCQ} is essentially stable (39.34% → 38.95%). In contrast, for RFT-QA, removing the prior prompt improves Acc_{QA}^{LLM} (25.22% → 29.17%), while also substantially boosting list performance (3.59% → 41.92% Acc_{List}^{LLM}).

No CoT prompt When removing the CoT instruction, effects are mixed. For RFT-MCQ, Acc_{MCQ} increased very slightly from 39.34% → 39.80%. For RFT-QA, QA performance is similar (25.22% → 24.78%), but list accuracy improves sharply (3.59% → 49.93% Acc_{List}^{LLM}). For RFT-List-Acc, list accuracy rises from 56.61% to 67.08% and QA accuracy improves (19.01% → 44.46%), though MCQ performance decreases (22.40% → 12.97%). Similarly, for RFT-List-MRR, Acc_{QA}^{LLM} improves from 21.90% to 46.21% and list accuracy grows slightly (61.60% → 63.00%), while MCQ drops (18.23% → 10.41%). We conjecture that without the thinking template, the model achieves higher accuracy due to the optimization objective in both QA and list answer formats evaluation (since QA is a special case of list), but at the cost of robustness in other formats.

Another notable side effect is that under the List prompt, the average list length increases dramatically (5.94 → 615.17), as the model tends to repeat sets of results. We observe similar behaviors in other models trained with the same reward functions (Section 4.3.2). While removing prior prompt or CoT suggests improved performance, a key trade-off is the loss of the <think>/</think> structure, which is important for certain test-time scaling techniques such as budget forcing (Mennighoff et al., 2025) or thinking interventions (Wu et al., 2025).

1830 I.5.2 INITIAL MODELS

1831
1832 Tables 7 and 8 report performance and list metrics for the experiments in Section 4.3.2. We observe
1833 that model family influences response length after RFT, broadly mirroring zero-shot response-length
1834 trends. For example, SFT-trained reasoning models retain high average token counts after RFT.
1835 For other families, however, response length does not necessarily correlate with performance (as
previously discussed) and varies without a consistent trend.

1836 Table 8: Average response length (mean \pm standard deviation) for MCQ, QA, and list-based answer
 1837 formats across benchmarks and metrics related to the ranked list answer format from the generated
 1838 evaluation responses for the initial model ablation study.

		MCQ	QA	List	CP	LL	VLL
RFT-MCQ							
	Qwen2.5 7B Instruct	204 \pm 106	167 \pm 65	33 \pm 123	1.45	2.29	2.29
	Qwen2.5 3B Instruct	197 \pm 213	190 \pm 159	179 \pm 94	1.35	1.94	1.94
	Qwen3 4B Instruct	845 \pm 613	554 \pm 509	453 \pm 599	1.44	3.40	3.46
	OpenThinker3 7B	1314 \pm 666	1376 \pm 1098	1042 \pm 953	1.38	2.58	2.62
	m1 7B 23K	1395 \pm 748	1105 \pm 703	1091 \pm 901	1.34	2.49	2.50
	AlphaMed 7B	342 \pm 286	309 \pm 663	278 \pm 608	1.08	0.60	1.45
RFT-QA							
	Qwen2.5 7B Instruct	296 \pm 302	291 \pm 247	46 \pm 91	1.67	0.29	3.01
	Qwen2.5 3B Instruct	264 \pm 306	301 \pm 298	327 \pm 640	1.78	4.16	4.19
	Qwen3 4B Instruct	946 \pm 874	673 \pm 736	524 \pm 536	1.48	3.64	3.64
	OpenThinker3 7B	1270 \pm 583	1088 \pm 562	791 \pm 467	1.41	2.96	2.96
	m1 7B 23K	1183 \pm 794	800 \pm 453	749 \pm 936	1.32	3.20	3.20
	AlphaMed 7B	256 \pm 142	246 \pm 337	220 \pm 237	1.19	1.55	1.57
RFT-List-Acc							
	Qwen2.5 7B Instruct	208 \pm 231	195 \pm 433	165 \pm 177	2.07	5.94	5.94
	Qwen2.5 3B Instruct	315 \pm 188	217 \pm 195	7881 \pm 1522	4.47	807.73	808.08
	Qwen3 4B Instruct	817 \pm 690	510 \pm 524	7929 \pm 1346	8.65	800.19	801.19
	OpenThinker3 7B	1519 \pm 889	1221 \pm 804	8087 \pm 826	5.95	828.88	831.05
	m1 7B 23K	1478 \pm 1208	1148 \pm 1211	8158 \pm 425	7.40	772.45	773.23
	AlphaMed 7B	256 \pm 114	202 \pm 323	3419 \pm 3782	1.75	4.82	4.83
RFT-List-MRR							
	Qwen2.5 7B Instruct	174 \pm 110	157 \pm 207	319 \pm 1154	2.11	16.97	16.97
	Qwen2.5 3B Instruct	334 \pm 236	240 \pm 198	7182 \pm 2601	4.04	733.87	734.94
	Qwen3 4B Instruct	752 \pm 787	537 \pm 787	52 \pm 94	2.17	5.79	5.79
	OpenThinker3 7B	1476 \pm 883	1195 \pm 817	8048 \pm 984	5.87	811.56	814.41
	m1 7B 23K	1543 \pm 1039	1249 \pm 1204	7679 \pm 1895	5.55	717.71	719.24
	AlphaMed 7B	277 \pm 118	198 \pm 285	218 \pm 95	1.66	5.48	5.48

J TRAINING DYNAMICS

We examine training dynamics using two metrics: (1) **reward progression** and (2) **response length trends**. These metrics allow us to holistically observe how changes in factors such as reward function, model family, or answer format affect the training process. The list results are presented and discussed alongside quantitative performance metrics for the main experiments, where applicable.

J.1 MAIN EXPERIMENT

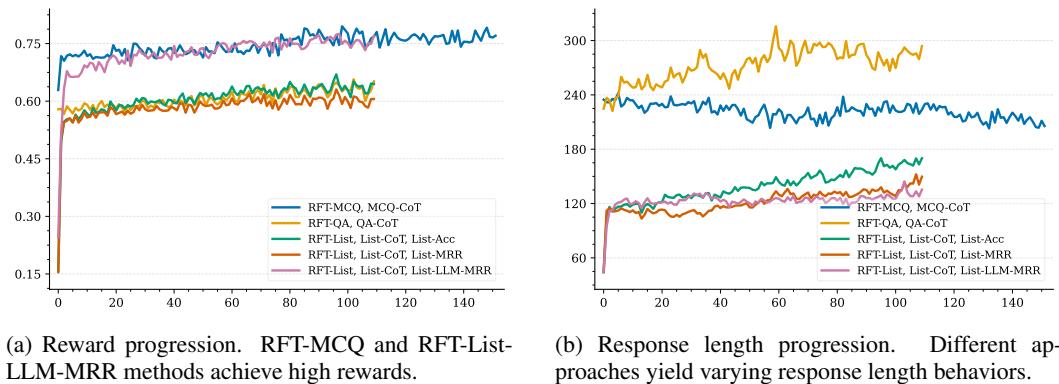


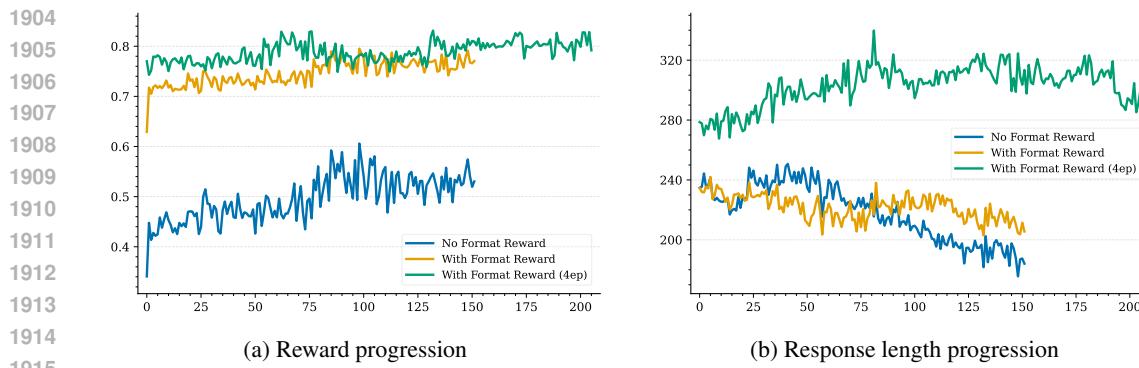
Figure 15: Training dynamics across different answer formats and reward functions.

This section provides a discussion on training dynamics of the models from Section 4.2. Training dynamics is provided in Figure 15. All RFT models share similar training dynamics with some differences. Among the RFT-List models, different reward types also lead to slightly different training dynamics, with the exception of RFT-List-Judge-MRR. For reward progression, most models exhibit a common trend: an initial low reward followed by a sharp increase, reflecting behavior aimed

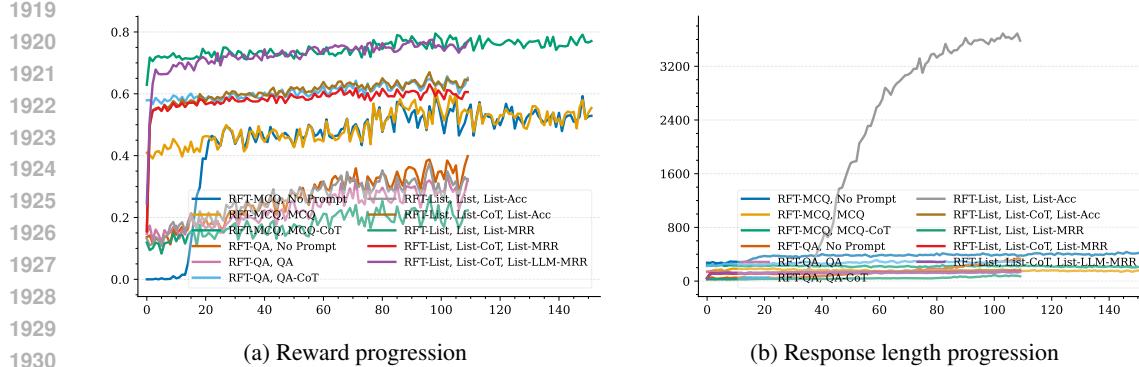
1890 at optimizing the format reward. The reward then continues to increase gradually over the course
 1891 of training. We also observe that RFT-MCQ and RFT-List-Judge-MRR achieve higher rewards than
 1892 the other models, suggesting that these models are able to score correct answers more consistently
 1893 under their respective reward types. QA is more challenging due to its reliance on exact match
 1894 rewards, similar to the list format. However, the final reward obtained during training does not reli-
 1895 ably predict final performance (Pearson $r = -0.267$, $p = 0.0671$, across all our RFT models in all
 1896 experiments).

1897 For response length progression, most models follow a pattern similar to reward progression: starting
 1898 with short responses that gradually increase in length. An exception is RFT-MCQ, which consis-
 1899 tently produces longer responses than the other models, and RFT-QA, which generates responses
 1900 that are longer than those from most other models.

1901 J.2 FACTORS AFFECTING RFT



1914 Figure 16: Training dynamics comparison between using and not using format reward, as well as
 1915 the effect of extending training from 2 to 4 epochs.



1916 Figure 17: Training dynamics comparison between different types of prior prompts and the case
 1917 without any prior prompt.

1918 Training dynamics are illustrated in Figures 16, 17 and 18b, which correspond respectively to exper-
 1919 iments on removing the format reward, extending training, altering or removing the prior prompt,
 1920 and changing the LLM judge.

1921 Removing the format reward does not substantially affect final model performance or list behaviors.
 1922 The primary differences lie in training dynamics: models without a format reward exhibit a lower
 1923 reward range during training, as they must focus exclusively on accuracy without a steady signal
 1924 of format reward. In addition, response length shows a slightly more pronounced decreasing trend
 1925 compared to models trained with the format reward.

1926 Extending training from two to four epochs does not substantially improve performance, except for
 1927 a tendency toward longer responses. Interestingly, the run with longer training begins with a higher
 1928 reward.

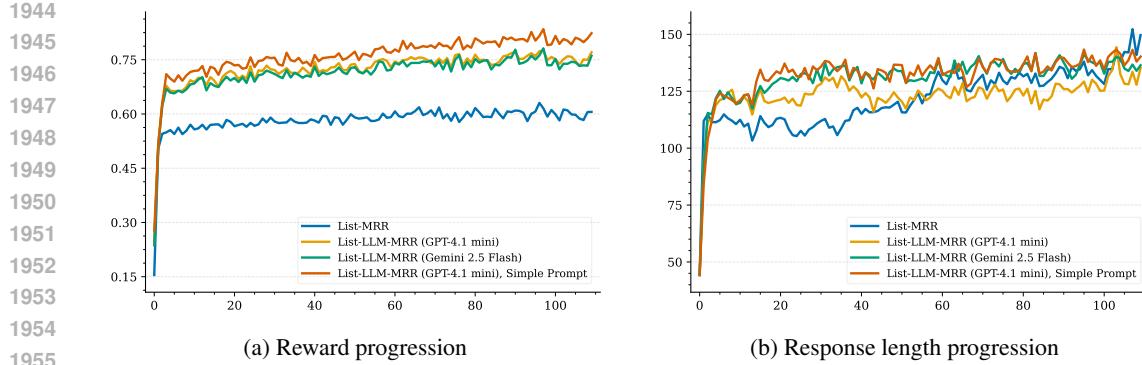


Figure 18: Training dynamics comparison across different Judge-MRR settings: GPT-4.1-mini and Gemini with the standard judge prompt, and GPT-4.1-mini with a simplified judge prompt.

initial reward and response length. Although, the overall training dynamics remain similar to the shorter run.

Models trained without prior prompts show lower initial rewards but experience a sharper increase later, while response length remains relatively stable throughout training.

J.3 INITIAL MODELS

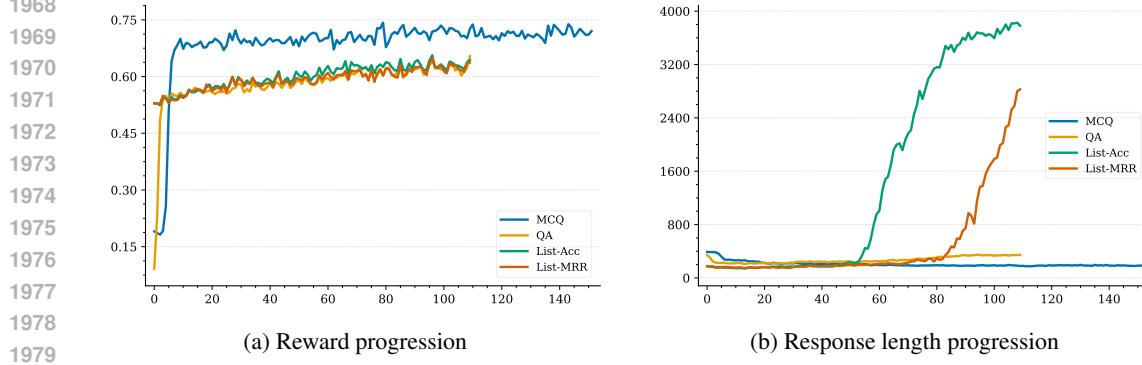


Figure 19: Training dynamics for Qwen2.5 3B Instruct.

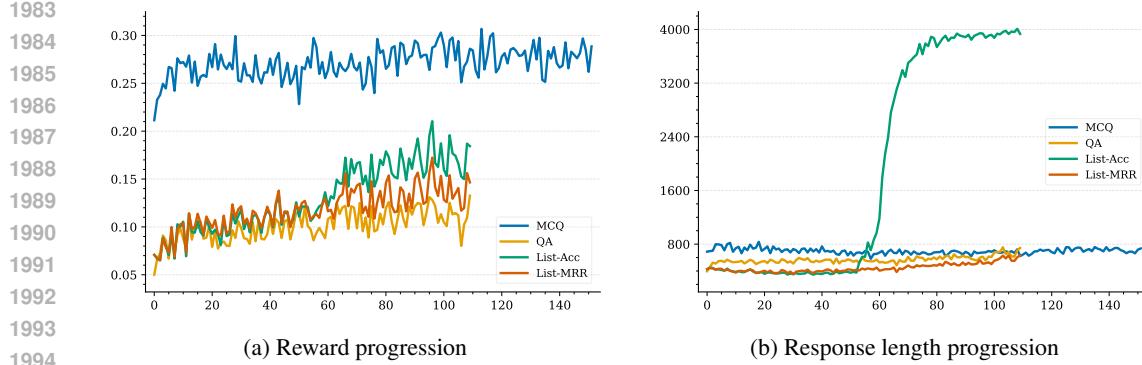


Figure 20: Training dynamics for Qwen3 4B.

Training dynamic of different initial models are illustrated in Figures 19 to 23.

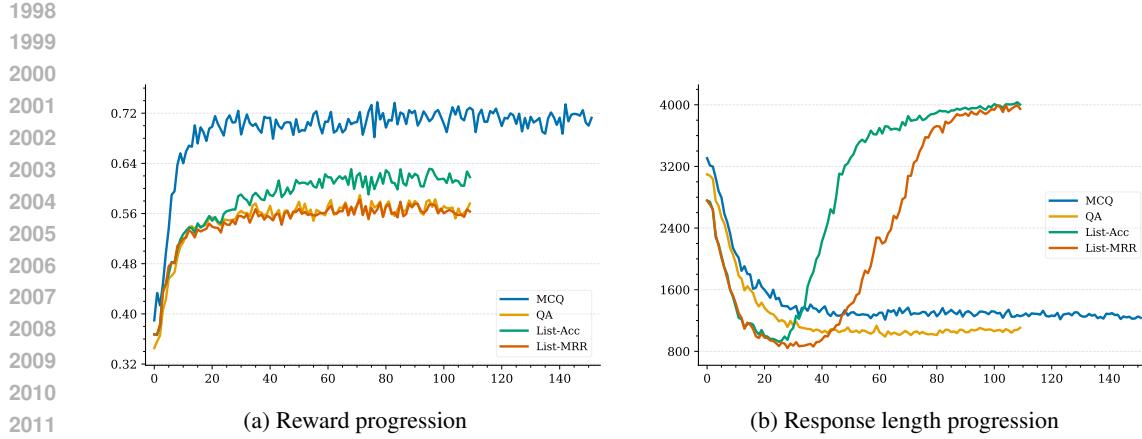


Figure 21: Training dynamics for OpenThinker 7B.

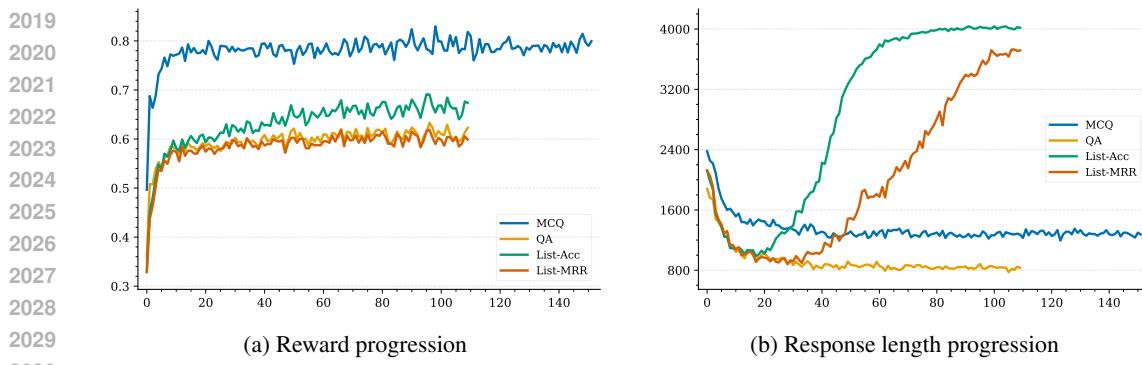


Figure 22: Training dynamics for m1 7B 23k.

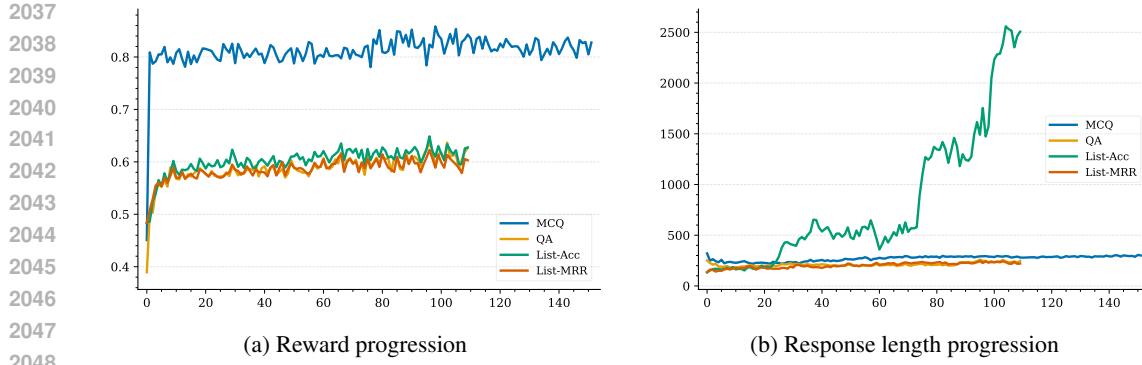


Figure 23: Training dynamics for AlphaMed 7B.

Table 9: Performance results of the ablation study on mixed datasets.

	MCQ		QA		List		
	Acc _{MCQ}	Acc _{QA}	Acc _{QA} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}
RFT-MCQ	39.34	9.96	46.33	9.89	40.06	8.16	33.00
RFT-QA	36.80	1.04	25.22	0.62	3.59	0.48	2.82
RFT-List-Acc	22.40	4.28	19.01	22.11	56.61	16.17	40.26
RFT-List-MRR	18.23	4.43	21.90	20.96	61.60	15.83	44.89
RFT-MCQ+QA	12.81	3.40	31.96	6.39	41.18	5.56	34.24
RFT-MCQ+List-Acc	39.70	11.24	46.89	21.47	61.52	15.10	41.93
RFT-MCQ+List-MRR	40.01	11.82	47.91	18.99	57.34	14.89	43.80

Table 10: Average response length (mean \pm standard deviation) for MCQ, QA, and list-based answer formats across benchmarks and metrics related to the ranked list answer format from the generated evaluation responses for the mixed datasets ablation study.

	MCQ	QA	List	CP	LL	VLL
RFT-MCQ	204 \pm 106	167 \pm 65	33 \pm 123	1.45	2.29	2.29
RFT-QA	296 \pm 302	291 \pm 247	46 \pm 91	1.67	0.29	3.01
RFT-List-Acc	208 \pm 231	195 \pm 433	165 \pm 177	2.07	5.94	5.94
RFT-List-MRR	174 \pm 110	157 \pm 207	319 \pm 1154	2.11	16.97	16.97
RFT-MCQ+QA	233 \pm 170	209 \pm 130	116 \pm 190	1.48	2.78	2.78
RFT-MCQ+List-Acc	246 \pm 209	201 \pm 173	193 \pm 541	2.45	10.34	10.35
RFT-MCQ+List-MRR	210 \pm 202	171 \pm 276	134 \pm 73	1.78	4.92	4.92

K RFT WITH MIXED ANSWER FORMAT DATASETS

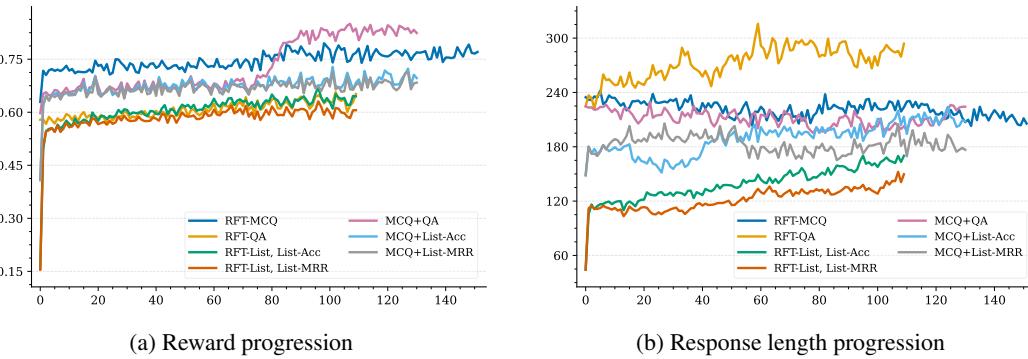


Figure 24: Training dynamics comparison across different dataset types: MCQ-only, QA-only, List-only, MCQ+QA, and MCQ+List.

To evaluate whether combining different answer formats can improve RFT, we construct mixed datasets by merging MCQ and QA data with appropriate prior prompts. Since mixing effectively doubles the dataset size, we train for one epoch instead of two to maintain a comparable number of optimization steps with the main experiments. Records are shuffled randomly, and rewards are computed according to the record type. Performance and list results are shown in Tables 9 and 10, and training dynamics are presented in Figure 24.

When training with a mixed dataset of MCQ and List using the MRR_{List} reward function, we observe the best overall performance on MCQ and QA compared to training on either dataset alone. However, this setting does not reach the strongest performance on the a ranked-list answer formats, where training exclusively on list data remains superior. A similar pattern holds for MCQ and List under the Acc_{List} reward.

In contrast, mixing MCQ and QA yields weaker results. While QA performance improves relative to training with QA alone, MCQ and list-format performance degrade noticeably. This outcome suggests that the mixture introduces instability, likely because QA is a sparse-reward task, making the overall reward signal less reliable when combined with denser MCQ data.

Table 11: Performance results of the ablation study on length penalty.

	MCQ		QA		List		
	Acc _{MCQ}	Acc _{QA}	Acc _{QA} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}
Qwen2.5 3B Instruct + LP with $\lambda = 0.3$	32.16 31.16	7.25 6.32	39.29 30.05	20.66 9.91	59.82 26.65	12.08 9.60	34.68 25.06
OpenThinker3 7B + LP with $\lambda = 0.3$	34.23 34.36	5.75 5.78	41.42 41.79	20.05 9.22	56.98 26.86	11.03 8.94	31.93 25.01
m1 7B 23K + LP with $\lambda = 0.3$	44.04 43.70	8.96 8.43	43.92 43.51	26.48 13.16	66.34 33.23	15.52 12.81	37.11 32.16
AlphaMed 7B + LP with $\lambda = 0.3$	38.35 21.82	5.82 9.65	25.38 38.60	17.86 13.48	57.29 33.55	14.32 13.25	44.16 31.94

Table 12: Average response length (mean \pm standard deviation) for MCQ, QA, and list-based answer formats across benchmarks and metrics related to the ranked list answer format from the generated evaluation responses for the length penalty ablation study.

	MCQ	QA	List	CP	LL	VLL
Qwen2.5 3B Instruct + LP with $\lambda = 0.3$	315 ± 188 230 ± 188	217 ± 195 178 ± 106	7881 ± 1522 131 ± 107	4.47 1.13	807.73 1.34	808.08 1.35
OpenThinker3 7B + LP with $\lambda = 0.3$	1519 ± 889 1288 ± 671	1221 ± 804 1077 ± 585	8087 ± 826 720 ± 479	5.95 1.15	828.88 1.41	831.05 1.41
m1 7B 23K + LP with $\lambda = 0.3$	1478 ± 1208 1370 ± 914	1148 ± 1211 890 ± 538	8158 ± 425 742 ± 450	7.40 1.09	772.45 1.22	773.23 1.22
AlphaMed 7B + LP with $\lambda = 0.3$	256 ± 114 213 ± 86	202 ± 323 174 ± 217	3419 ± 3782 151 ± 126	1.75 1.11	4.82 1.31	4.83 1.31

Across all experiments, we find that mixing answer formats in the same dataset is not particularly effective. None of the models trained on mixed datasets produce excessively long lists (e.g., >100 items), but performance trade-offs prevent mixed training from outperforming single-answer-format training in most cases.

L LIST REWARD HACKING AND MITIGATION WITH LENGTH PENALTY

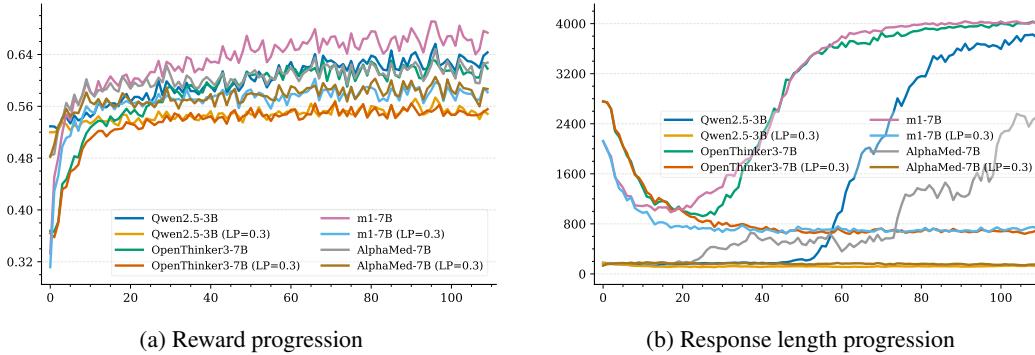


Figure 25: Training dynamics comparison for models before and after applied the length penalty.

As shown in Sections 4.3.1 and 4.3.2, some models exhibit a tendency to generate excessively long lists (sometimes exceeding 100 items) when trained with list-based reward functions. While training on mixed-format datasets (Appendix K) can partially mitigate this issue, we propose an alternative approach: modifying the list reward functions themselves by introducing a **length penalty term**.

Both Acc_{List} and MRR_{List} reward functions incentivize correctness but do not discourage unnecessarily long outputs. To address this, we introduce a length penalty term (*LP*) that scales the reward according to the number of items in the list. Let L denote the length of the generated list and λ the

Table 13: Performance results of the ablation study on length penalty.

	MCQ		QA		List		
	Acc _{MCQ}	Acc _{QA}	Acc _{QA} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}
LP=0	12.93	8.23	34.55	19.89	54.95	15.78	41.59
LP=0.1	37.73	11.85	47.23	13.90	36.02	13.46	33.40
LP=0.3	33.52	11.54	48.30	12.64	28.64	12.34	27.71
LP=0.5	37.78	11.24	46.25	11.86	26.02	11.86	26.01
LP=0.7	34.71	11.56	47.35	11.68	27.39	11.68	27.34
LP=0.9	35.47	10.34	46.08	11.83	27.07	11.83	27.05

Table 14: Average response length (mean \pm standard deviation) for MCQ, QA, and list-based answer formats across benchmarks and metrics related to the ranked list answer format from the generated evaluation responses for the length penalty ablation study.

	MCQ	QA	List	CP	LL	VLL
LP=0	170 \pm 141	166 \pm 296	132 \pm 141	1.78	4.71	4.71
LP=0.1	176 \pm 207	170 \pm 423	127 \pm 174	1.18	1.52	1.52
LP=0.3	160 \pm 272	164 \pm 372	11 \pm 6	1.08	1.16	1.16
LP=0.5	157 \pm 76	139 \pm 100	10 \pm 3	1.00	1.01	1.01
LP=0.7	161 \pm 186	166 \pm 362	11 \pm 4	1.01	1.05	1.05
LP=0.9	60 \pm 117	64 \pm 80	11 \pm 4	1.00	1.03	1.03

penalty coefficient. The penalty term is defined as:

$$LP = \max(0, 1 - \lambda \cdot (L - 1)).$$

Since LP is orthogonal to existing reward functions, it can be applied to both Acc_{List} and MRR_{List} by first computing the correctness reward and then scaling it by LP . In both cases, the penalty encourages concise and precise outputs rather than exhaustive or repetitive enumerations. This introduces an additional optimization constraint: the model must both produce the correct answer and avoid generating unnecessarily long lists. While we adopt the simplest linear form of length penalty here, future work may explore more flexible variants that allow stronger or weaker tolerance for longer outputs.

Effectiveness of LP in mitigating large lists We evaluate the effectiveness of the length penalty by applying it to four models that previously exhibited excessively long lists when trained with the Acc_{List} reward function: Qwen2.5 3B Instruct, OpenThinker3 7B, m1 7B 23k, and AlphaMed 7B. We set $\lambda = 0.3$ as an arbitrary but fixed choice. Results in Table 12 show that the length penalty successfully reduces the large-list behavior across all models. Training dynamics in Figure 25 further demonstrate that response length remains more stable over time, with similar reward progression trends compared to the unpenalized setup, albeit at slightly lower reward values. We also observe that the gap between Acc_{List} and MRR_{List} narrows in the evaluation results (Table 11), as the length penalty encourages correct items to appear in higher positions, thereby reducing the average correct position (Table 12).

Trade-offs of LP between list length and performance Although the length penalty effectively reduces uncontrolled list growth, it introduces trade-offs. Across models, we observe moderate reductions in MCQ and QA performance, and more substantial drops in a ranked-list answer format performance, where the penalty directly applies. This highlights the inherent trade-off between controlling undesirable behaviors and maximizing performance. As noted, our penalty term represents a simple first step and could be refined to better balance this trade-off. Larger models, which we do not investigate due to computational constraints, may also be less sensitive to such penalties. Additional experiments on varying λ are presented in Appendix M.

M LENGTH PENALTY HYPERPARAMETERS

In this section, we examine how different values of λ affect both performance and list length. We select $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ to cover a range from mild to strong penalization. To demonstrate the generality of the LP formulation, we study its application to the MRR_{List} reward function using

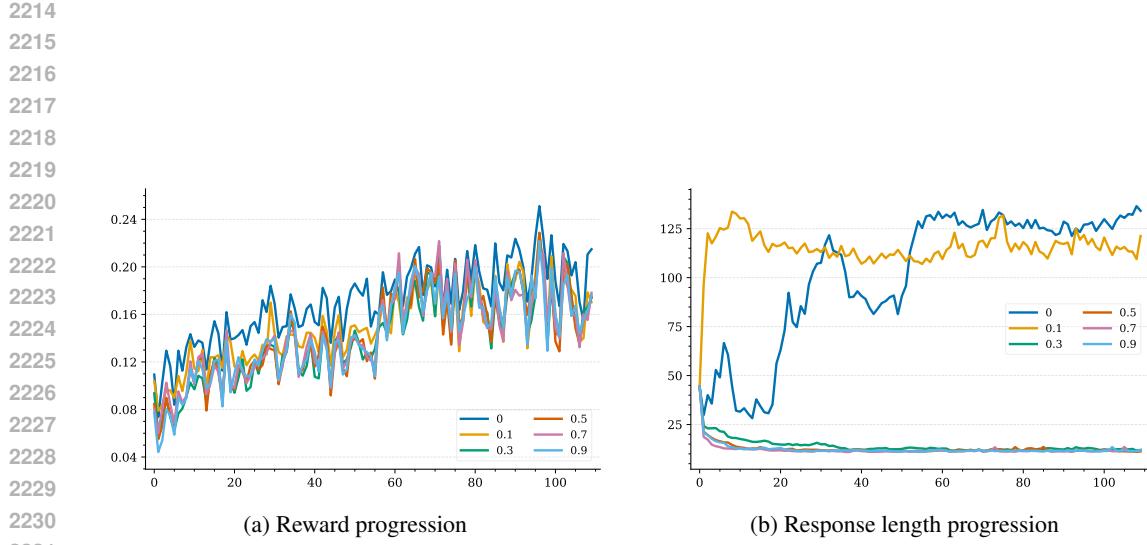
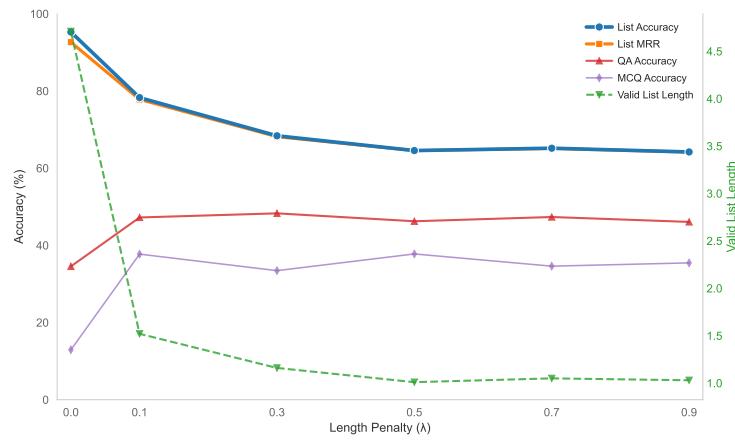


Figure 26: Training dynamics comparison for length penalty ablation.

Figure 27: Impact of length penalty on model performance. Performance trends across different task types, where List tasks demonstrate highest baseline performance but steep degradation, QA tasks exhibit optimal performance at $\lambda = 0.1\text{--}0.3$, and MCQ tasks show consistent improvement with positive length penalty.

2263
2264
2265
2266
2267

2268 Table 15: Results table for prompting experiments evaluated on MCQ benchmarks using both direct
 2269 **MCQ** and **MCQ-CoT** prompting strategies from Section 3. MXQA refers to MedXpertQA (text),
 2270 and MLUP-H refers to MMLU Pro (Health).

	MedQA	MedMCQA	MXQA	MLUP-H	Average
<i>Proprietary Reasoning Models</i>					
Gemini 2.5 Flash Lite	82.01	18.80	20.82	72.25	48.47
+CoT	50.98	9.30	11.27	30.81	25.59
Gemini 2.5 Flash	92.22	18.05	37.39	73.11	55.19
+CoT	61.67	7.51	29.92	42.67	35.44
Gemini 2.5 Pro	94.34	18.78	44.94	76.65	58.68
+CoT	94.82	16.72	45.76	74.94	58.06
<i>Proprietary Large Language Models</i>					
GPT-4.1 Mini	90.49	18.91	31.59	77.87	54.72
+CoT	72.74	18.10	29.59	70.42	47.71
<i>Open-weight Large Language Models</i>					
Qwen2.5 3B Instruct	48.08	24.11	8.98	37.29	29.62
+CoT	52.24	23.63	11.84	43.03	32.68
Qwen2.5 7B Instruct	17.99	10.20	6.08	19.44	13.43
+CoT	49.10	17.58	12.33	45.97	31.24
Qwen2.5 14B Instruct	59.23	19.51	10.12	54.65	35.88
+CoT	54.91	19.56	10.20	60.64	36.33
Qwen3 4B Instruct 2507	73.21	19.53	16.16	66.38	43.82
+CoT	68.34	17.48	16.29	62.35	41.12
Gemma 3 4B IT	49.80	21.04	12.00	38.88	30.43
+CoT	46.27	16.57	11.59	38.14	28.14
<i>Open-weight Medical LLMs</i>					
MedGemma 4B IT	63.79	21.48	14.69	48.41	37.09
+CoT	54.83	16.55	13.76	36.92	30.52
MedGemma 27B IT	81.46	20.21	22.45	71.76	48.97
+CoT	65.44	10.81	16.90	37.78	32.73
<i>Open-weight Reasoning Models</i>					
OpenThinker3 7B	48.15	18.07	5.55	38.51	27.57
+CoT	3.46	3.15	2.41	5.99	3.75
<i>Open-weight Medical Reasoning Models</i>					
HuatuoGPT o1 7B	17.67	9.19	9.06	35.09	17.75
+CoT	6.28	6.47	5.47	22.62	10.21
m1 7B 23K	64.10	20.65	15.92	56.36	39.26
+CoT	53.73	15.43	12.12	44.25	31.38
AlphaMed 7B Instruct RL	69.91	16.49	16.24	59.41	40.51
+CoT	60.41	14.54	16.73	58.19	37.47

2303 Qwen2.5 7B Instruct. Performance and list results are presented in Tables 13 and 14, and training
 2304 dynamics are shown in Figure 26.

2305 Introducing the length penalty consistently reduces the length of generated lists across λ values,
 2306 confirming its effectiveness in controlling overly long outputs. However, this comes at the cost
 2307 of reduced performance, as models become more reluctant to produce longer lists. The trade-offs
 2308 discussed in Appendix L are observed across all choices of λ . Interestingly, applying the length
 2309 penalty also improves performance on single-answer tasks such as MCQ and QA. This suggests that
 2310 constraining the model to produce shorter, more focused outputs may indirectly benefit tasks where
 2311 concise responses are essential.

2313 N FULL RESULTS

2315 This section provides a detailed breakdown of results for each benchmark based on experiments
 2316 reported across papers.

2318 N.1 PROMPTING

2319 Tables 15 to 17 present the non-aggregated versions of the aggregated results shown in the main
 2320 body of this study (Table 1 in Section 3). While, Tables 18 to 20 show average response length

Table 16: Results table for prompting experiments evaluated on open-ended benchmarks using both direct **QA** and **QA-CoT** prompting strategies from Section 3. MXQA refers to MedXpertQA (text), and MLUP-H refers to MMLU Pro (Health).

	MedQA		MedMCQA		MXQA		MLUP-H		Average	
	Acc _{QA}	Acc _{LLM} ^{QA}	Acc _{QA}	Acc _{LLM} ^{QA}	Acc _{QA}	Acc _{LLM} ^{QA}	Acc _{QA}	Acc _{LLM} ^{QA}	Acc _{QA}	Acc _{LLM} ^{QA}
	Proprietary Reasoning Models									
Gemini 2.5 Flash Lite	19.63	53.69	10.00	45.69	6.33	41.18	13.45	54.21	12.35	48.69
+CoT	18.17	48.34	8.81	37.75	6.14	34.66	12.23	48.91	11.34	42.42
Gemini 2.5 Flash	17.03	50.45	9.91	43.76	8.49	42.38	12.36	47.83	11.95	46.10
+CoT	17.84	49.88	9.86	41.56	6.90	40.27	11.14	45.92	11.44	44.41
Gemini 2.5 Pro	17.52	54.74	8.58	44.82	7.00	44.53	10.46	52.72	10.89	49.20
+CoT	16.46	52.55	7.75	45.83	6.57	42.86	10.46	51.09	10.31	48.08
Proprietary Large Language Models										
GPT-4.1 Mini	13.87	52.31	8.90	45.64	5.75	39.98	8.70	50.14	9.30	47.02
+CoT	16.22	52.72	9.40	46.24	6.66	42.67	10.46	54.35	10.68	49.00
Open-weight Large Language Models										
Qwen2.5 3B Instruct	7.38	36.25	4.72	36.06	2.49	33.22	8.15	37.09	5.68	35.66
+CoT	8.27	41.28	4.95	40.46	2.73	34.85	7.61	44.16	5.89	40.19
Qwen2.5 7B Instruct	11.76	45.90	6.42	43.35	3.50	35.14	9.92	48.37	7.90	43.19
+CoT	12.08	46.80	6.47	41.19	3.88	35.67	11.41	49.73	8.46	43.35
Qwen2.5 14B Instruct	15.25	49.64	9.13	44.68	4.89	38.64	12.36	48.37	10.41	45.33
+CoT	15.33	53.20	9.04	47.11	4.65	39.41	11.41	51.49	10.11	47.80
Qwen3 4B Instruct 2507	15.49	51.50	8.30	45.00	5.18	40.36	12.91	52.04	10.47	47.22
+CoT	16.06	50.45	7.94	40.50	4.55	39.65	11.55	48.10	10.02	44.68
Gemma 3 4B IT	8.19	36.58	5.60	40.28	3.07	30.92	8.02	38.72	6.22	36.62
+CoT	9.73	40.31	5.28	40.92	2.88	35.67	9.10	43.75	6.75	40.16
Open-weight Medical LLMs										
MedGemma 4B IT	12.81	46.55	6.33	42.89	3.02	35.76	9.24	47.55	7.85	43.19
+CoT	12.33	47.61	6.97	41.93	3.31	34.80	10.46	47.01	8.27	42.84
MedGemma 27B IT	20.36	54.66	9.91	44.04	7.00	40.22	13.86	51.63	12.78	47.64
+CoT	15.82	45.01	9.04	40.28	5.32	32.36	12.23	46.20	10.60	40.96
Open-weight Reasoning Models										
OpenThinker3 7B	6.33	32.77	4.72	31.28	1.87	24.35	8.15	35.73	5.27	31.03
+CoT	5.92	33.41	4.36	31.56	1.97	23.30	7.20	34.51	4.86	30.70
Open-weight Medical Reasoning Models										
HuatuogPT o1 7B	1.30	3.24	0.87	4.04	0.58	2.88	1.63	5.16	1.10	3.83
+CoT	0.65	2.43	0.37	1.79	0.19	1.87	1.09	2.72	0.57	2.20
m1 7B 23K	12.33	43.15	6.38	38.99	3.74	32.17	8.15	41.58	7.65	38.97
+CoT	12.00	43.80	6.10	35.78	2.83	28.04	9.92	41.44	7.71	37.26
AlphaMed 7B Instruct RL	0.41	5.52	1.10	13.53	0.00	7.38	1.77	11.41	0.82	9.46
+CoT	8.76	25.22	5.28	28.30	2.35	17.40	6.39	26.90	5.70	24.46

Table 17: Results table for prompting experiments evaluated on open-ended benchmarks using both direct **List** and **List-CoT** prompting strategies from Section 3.

Model	MedQA			MedMCQA			MedXpertQA (Text)			MMLU Pro (Health)			Average							
	Acc _{List}	Acc _{LLM} ^{List}	MRR _{List}	Acc _{List}	Acc _{LLM} ^{List}	MRR _{List}	Acc _{List}	Acc _{LLM} ^{List}	MRR _{List}	Acc _{List}	Acc _{LLM} ^{List}	MRR _{List}	Acc _{List}	Acc _{LLM} ^{List}	MRR _{List}					
	Proprietary Reasoning Models																			
Gemini 2.5 Flash Lite	21.74	63.83	19.96	57.07	13.17	41.93	11.54	34.71	7.92	46.60	7.92	35.93	14.54	64.27	13.94	58.74	14.84	54.16	13.34	46.61
+CoT	10.14	25.22	9.70	24.11	9.17	24.54	8.72	22.92	3.84	13.52	3.46	11.97	9.24	34.92	9.01	33.42	8.10	24.55	7.72	23.10
Gemini 2.5 Flash	25.79	74.37	23.18	65.24	15.72	44.82	12.06	37.38	13.47	60.93	11.46	49.40	16.98	72.01	15.35	62.80	17.49	63.03	15.51	53.70
+CoT	9.04	27.01	9.06	26.31	10.26	24.81	7.69	23.66	10.37	62.05	8.07	36.14	7.44	37.55	7.72	37.23	8.04	34.38	7.91	34.38
Gemini 2.5 Pro	25.71	78.51	23.86	90.88	15.41	38.39	13.11	38.58	15.77	67.26	12.98	55.93	10.65	76.49	15.81	67.92	18.44	67.66	16.44	58.33
+CoT	23.28	76.24	21.24	68.21	14.45	46.61	12.17	37.21	13.71	64.96	11.21	53.62	14.54	73.10	13.32	63.71	16.50	65.23	14.48	55.69
Proprietary Large Language Models																				
GPT-4.1 Mini	18.98	73.15	17.82	67.10	11.88	47.89	10.29	39.48	9.40	56.09	7.39	45.96	10.87	70.11	10.20	63.60	12.78	61.81	11.42	54.04
+CoT	17.84	75.59	16.66	69.97	12.34	50.09	10.36	40.55	10.39	59.59	8.39	48.96	10.87	74.05	10.12	67.41	12.85	64.83	11.38	56.63
Open-weight Large Language Models																				
Qwen2.5 3B Instruct	13.38	46.55	10.45	35.76	7.57	29.72	6.08	23.26	4.89	29.67	3.55	20.28	10.87	50.41	9.32	39.93	9.18	39.09	7.35	29.80
+CoT	11.44	25.20	10.63	31.70	7.40	23.24	6.24	20.41	3.54	31.04	3.44	41.37	8.23	37.23	7.52	29.63	10.71	26.59		
Qwen2.5 7B Instruct	14.76	48.91	11.97	40.26	7.43	37.84	6.78	24.29	5.90	30.87	4.89	23.44	10.45	53.40	11.61	45.35	10.38	40.26	8.81	33.46
+CoT	16.30	54.66	13.40	44.62	10.00	40.78	7.90	31.20	7.53	38.40	5.11	27.36	10.33	60.46	9.06	50.84	11.04	48.58	8.87	38.50
Qwen2.5 14B Instruct	18.33	60.05	15.83	50.33	13.21	42.84	10.71	34.42	8.01	41.80	6.15	30.67	13.32	66.30	12.09	58.36	13.09	52.74	11.20	43.44
+CoT	19.22	59.61	13.04	49.33	13.73	40.32	8.32	31.89	8.86	42.14	4.74	29.21	11.14	66.40	8.57	58.83	11.39	53.85	8.96	40.77
Qwen3 4B Instruct 2507	16.95	60.31	16.76	42.60	10.78	39.22	7.29	33.72	7.52	40.85	5.10	35.69	10.82	58.15	11.80	51.82	12.39	49.89	10.87	41.54
+CoT	19.22	61.64	16.76	42.60	10.78	39.22	7.29	33.72	7.52	40.85	5.10	35.69	10.82	58.15	11.80	51.82	12.39	49.89	10.87	41.54
Gemma 3 4B IT	17.52	53.93	12.93	40.65	10.60	38.58	7.40	26.52	7.14	38.11	4.63	25.88	11.82	59.83	9.94	46.17	11.77	47.36	8.72	34.80
+CoT	12.81	45.26	10.17	36.10	7.25	27.75	5.61	20.44	5.56	28.04	3.90	20.56	11.55	47.28	9.07	37.90	9.29	37.08	7.19	28.75
Open-weight Medical LLMs																				
MedGemma 4B IT	18.09	60.58	13.62	46.27	13.81	44.54	9.07	29.71	8.29	42.43	5.11	27.58	13.18	65.08	10.52	51.55	13.34	53.16	9.58	38.78
+CoT	20.44	61.64	15.73	47.09	12.89	44.88	8.91	27.84	8.40	40.48	4.98	25.00	11.96	57.47	9.70	44.89	13.56	50.61	9.83	38.41
MedGemma 27B IT	21.98	63.50	19.30	55.73	13.21	38.76	11.31	32.65	9.16	40.65	7.22	31.15	13.45	59.51	13.29	43.44	14.45	50.60	12.75	43.24
+CoT	22.30	66.10	20.33	58.91	12.80	39.91	11.20	33.39	10.55	43.53	8.36	34.38	15.08	63.18	14.21	57.07	15.18	53.18	13.52	45.94
Open-weight Reasoning Models																				
OpenThinker3 7B	6.97	33.25	6.39	29.1																

2376

2377

2378

2379

2380

2381

2382

2383

2384

2385

2386

2387

Table 18: Response length (mean \pm standard deviation) for **MCQ** answer format prompting across benchmarks from Section 3.

2388

2389

2390

2391

	MedQA	MedMCQA	MXQA	MMLU Pro-H	Average
<i>Proprietary Reasoning Models</i>					
Gemini 2.5 Flash Lite	1257.47 \pm 3191.23	1119.54 \pm 3512.21	2628.94 \pm 4957.93	1335.20 \pm 3220.06	1585.29
+CoT	2818.61 \pm 4419.52	2139.62 \pm 4667.54	8736.47 \pm 7822.16	3745.60 \pm 5907.18	4360.07
Gemini 2.5 Flash	545.11 \pm 197.13	123.44 \pm 192.86	844.54 \pm 266.83	377.23 \pm 346.71	472.58
+CoT	1587.31 \pm 1154.53	1033.27 \pm 1025.54	2967.20 \pm 2043.39	1449.88 \pm 1243.58	1759.41
Gemini 2.5 Pro	490.35 \pm 138.99	248.95 \pm 127.13	627.53 \pm 270.05	334.02 \pm 217.40	425.21
+CoT	1465.06 \pm 266.79	954.33 \pm 272.60	1631.28 \pm 390.26	1252.48 \pm 375.49	1325.79
<i>Proprietary Large Language Models</i>					
GPT-4.1 Mini	415.21 \pm 150.39	155.87 \pm 75.99	539.88 \pm 224.97	302.50 \pm 171.92	353.37
+CoT	382.19 \pm 118.66	200.29 \pm 67.40	509.30 \pm 165.14	334.75 \pm 138.41	356.63
<i>Open-weight Large Language Models</i>					
Qwen2.5 3B Instruct	355.70 \pm 247.30	209.95 \pm 233.14	197.95 \pm 235.02	318.69 \pm 169.92	270.57
+CoT	368.60 \pm 328.27	263.84 \pm 257.55	431.82 \pm 501.80	381.64 \pm 133.37	361.48
Qwen2.5 7B Instruct	112.42 \pm 83.21	39.83 \pm 48.06	83.12 \pm 70.98	53.92 \pm 68.11	72.32
+CoT	1066.64 \pm 2476.43	1819.36 \pm 3206.03	3376.30 \pm 3906.91	3310.99 \pm 3829.30	2393.32
Qwen2.5 14B Instruct	169.53 \pm 81.65	89.49 \pm 65.35	149.53 \pm 68.53	108.68 \pm 72.01	129.31
+CoT	227.06 \pm 79.52	171.93 \pm 190.53	239.99 \pm 104.90	228.72 \pm 116.49	216.93
Qwen3 4B Instruct 2507	917.31 \pm 1093.89	364.78 \pm 640.04	1074.26 \pm 899.40	669.64 \pm 835.01	756.50
+CoT	1055.03 \pm 1189.65	536.58 \pm 793.92	1191.33 \pm 936.30	794.32 \pm 797.72	894.32
Gemma 3 4B IT	407.36 \pm 98.68	255.02 \pm 81.38	529.57 \pm 136.83	406.46 \pm 131.04	399.60
+CoT	403.89 \pm 782.49	273.21 \pm 607.06	425.57 \pm 544.08	424.09 \pm 740.47	381.69
<i>Open-weight Medical LLMs</i>					
MedGemma 4B IT	376.22 \pm 992.80	122.10 \pm 541.86	427.47 \pm 1056.55	263.80 \pm 646.20	297.40
+CoT	7657.11 \pm 1975.69	8060.86 \pm 1009.42	8188.78 \pm 155.49	7924.15 \pm 1423.73	7957.73
MedGemma 27B IT	1193.72 \pm 855.60	572.05 \pm 732.17	1585.79 \pm 1010.40	974.36 \pm 789.28	1081.48
+CoT	1419.63 \pm 826.18	852.78 \pm 783.07	2090.85 \pm 1027.10	1331.33 \pm 874.37	1423.65
<i>Open-weight Reasoning Models</i>					
OpenThinker3 7B	4386.88 \pm 2777.18	3672.76 \pm 2787.49	6164.33 \pm 2439.96	4933.49 \pm 2773.43	4789.36
+CoT	7538.26 \pm 1933.86	7284.63 \pm 2249.75	7139.14 \pm 2387.52	7399.08 \pm 1985.14	7340.28
<i>Open-weight Medical Reasoning Models</i>					
HuatuoGPT o1 7B	501.45 \pm 83.88	399.07 \pm 124.61	515.03 \pm 179.38	472.94 \pm 287.54	472.12
+CoT	524.25 \pm 315.82	410.86 \pm 189.21	535.78 \pm 238.44	479.55 \pm 96.68	487.61
m1 7B 23K	1423.43 \pm 2067.20	1117.32 \pm 1798.36	1958.03 \pm 2502.58	1811.57 \pm 2445.53	1577.59
+CoT	8185.63 \pm 195.05	8184.38 \pm 214.57	8188.19 \pm 135.88	8182.06 \pm 262.03	8185.07
AlphaMed 7B Instruct RL	336.71 \pm 252.38	197.61 \pm 155.88	362.00 \pm 210.89	346.32 \pm 129.34	310.66
+CoT	400.09 \pm 330.13	322.74 \pm 357.22	489.54 \pm 744.54	452.29 \pm 488.71	416.16

2419

2420

2421

2422

2423

2424

2425

2426

2427

2428

2429

2430

2431

2432

2433

2434

2435

2436

2437

2438

2439

2440

2441

Table 19: Response length (mean \pm standard deviation) for **QA** answer format prompting across benchmarks from Section 3.

2443

2444

2445

	MedQA	MedMCQA	MXQA	MMLU Pro-H	Average
<i>Proprietary Reasoning Models</i>					
Gemini 2.5 Flash Lite	539.03 \pm 713.66	366.96 \pm 1111.66	752.54 \pm 1854.21	386.84 \pm 306.99	511.34
+CoT	2578.72 \pm 3754.91	2193.28 \pm 4331.44	4146.62 \pm 5493.75	2481.86 \pm 4558.74	2850.12
Gemini 2.5 Flash	312.46 \pm 224.98	181.34 \pm 184.48	341.99 \pm 238.11	258.09 \pm 232.62	273.47
+CoT	1637.81 \pm 980.09	1206.37 \pm 885.25	2051.34 \pm 1222.96	1406.75 \pm 1055.64	1575.57
Gemini 2.5 Pro	499.26 \pm 286.12	363.56 \pm 278.51	545.75 \pm 290.95	428.83 \pm 298.50	459.35
+CoT	1555.26 \pm 463.10	1402.16 \pm 784.48	1655.74 \pm 479.10	1494.77 \pm 545.76	1526.98
<i>Proprietary Large Language Models</i>					
GPT-4.1 Mini	340.36 \pm 178.95	179.09 \pm 103.89	388.24 \pm 199.08	231.40 \pm 148.30	284.77
+CoT	320.84 \pm 119.89	238.50 \pm 85.35	355.43 \pm 134.23	253.05 \pm 102.18	291.95
<i>Open-weight Large Language Models</i>					
Qwen2.5 3B Instruct	246.31 \pm 343.36	170.46 \pm 268.18	245.44 \pm 279.15	193.74 \pm 123.33	213.99
+CoT	314.92 \pm 117.15	307.22 \pm 595.46	340.17 \pm 402.12	296.82 \pm 426.34	314.78
Qwen2.5 7B Instruct	171.16 \pm 251.37	108.60 \pm 89.28	170.38 \pm 108.71	133.57 \pm 101.85	145.93
+CoT	214.56 \pm 82.16	164.49 \pm 252.89	225.74 \pm 85.32	179.28 \pm 78.02	196.02
Qwen2.5 14B Instruct	160.09 \pm 244.07	118.92 \pm 78.74	156.92 \pm 92.65	122.62 \pm 78.94	139.64
+CoT	217.01 \pm 75.96	195.16 \pm 305.39	240.56 \pm 261.87	199.52 \pm 303.65	213.06
Qwen3 4B Instruct 2507	474.31 \pm 438.88	419.78 \pm 1037.35	563.43 \pm 679.78	341.24 \pm 485.06	449.69
+CoT	565.40 \pm 482.77	521.64 \pm 1000.22	624.52 \pm 600.58	441.24 \pm 500.05	538.20
Gemma 3 4B IT	480.07 \pm 201.84	339.02 \pm 188.74	504.22 \pm 198.02	405.77 \pm 210.79	432.27
+CoT	343.19 \pm 150.29	251.72 \pm 210.98	343.45 \pm 154.01	275.65 \pm 136.41	303.50
<i>Open-weight Medical LLMs</i>					
MedGemma 4B IT	166.35 \pm 345.02	108.60 \pm 528.33	183.20 \pm 381.67	150.28 \pm 531.34	152.11
+CoT	343.13 \pm 962.56	600.20 \pm 1810.89	526.34 \pm 1473.62	502.38 \pm 1534.13	493.01
MedGemma 27B IT	797.26 \pm 678.04	395.62 \pm 483.65	901.12 \pm 929.02	518.22 \pm 454.65	653.06
+CoT	1020.20 \pm 863.48	849.69 \pm 1055.90	1168.26 \pm 1020.78	826.97 \pm 854.14	966.28
<i>Open-weight Reasoning Models</i>					
OpenThinker3 7B	4372.27 \pm 2887.58	4053.37 \pm 2843.18	5123.57 \pm 2780.84	3843.64 \pm 2864.54	4348.21
+CoT	4483.61 \pm 2925.64	4330.32 \pm 2859.00	5184.47 \pm 2810.99	3801.95 \pm 2852.59	4450.09
<i>Open-weight Medical Reasoning Models</i>					
HuatuoGPT o1 7B	509.17 \pm 92.77	456.40 \pm 299.47	530.70 \pm 252.94	484.12 \pm 413.76	495.10
+CoT	532.43 \pm 323.88	455.65 \pm 84.25	543.79 \pm 303.99	471.90 \pm 96.36	500.94
m1 7B 23K	1628.53 \pm 2083.86	1376.54 \pm 1753.36	1909.36 \pm 2242.62	1254.38 \pm 1786.37	1542.20
+CoT	2355.00 \pm 2511.52	2068.92 \pm 2256.22	2795.45 \pm 2593.80	2091.34 \pm 2332.12	2327.68
AlphaMed 7B Instruct RL	244.48 \pm 406.33	186.20 \pm 312.77	235.77 \pm 272.32	198.74 \pm 108.44	216.30
+CoT	284.31 \pm 605.91	223.78 \pm 549.05	309.03 \pm 678.04	245.06 \pm 595.63	265.55

2473

2474

2475

2476

2477

2478

2479

2480

2481

2482

2483

2484

2485

2486

2487

2488

2489

2490

2491

2492

2493

2494

2495

Table 20: Response length (mean \pm standard deviation) for ranked-list answer format prompting across benchmarks from Section 3.

	MedQA	MedMCQA	MXQA	MMLU Pro-H	Average
<i>Proprietary Reasoning Models</i>					
Gemini 2.5 Flash Lite	98.61 \pm 744.16	88.90 \pm 997.32	135.87 \pm 1085.66	50.13 \pm 111.02	93.38
+CoT	7697.65 \pm 8361.72	5191.43 \pm 7473.53	10339.69 \pm 8598.17	6364.13 \pm 8108.09	7398.23
Gemini 2.5 Flash	34.67 \pm 28.84	21.87 \pm 24.46	39.85 \pm 41.35	31.26 \pm 38.97	31.91
+CoT	1090.39 \pm 793.30	820.04 \pm 755.68	1209.99 \pm 1023.00	962.94 \pm 854.95	1020.84
Gemini 2.5 Pro	41.62 \pm 45.35	26.98 \pm 41.48	44.23 \pm 45.51	41.07 \pm 88.56	38.47
+CoT	1073.70 \pm 254.87	885.01 \pm 275.27	1133.54 \pm 264.44	985.43 \pm 332.91	1019.42
<i>Proprietary Large Language Models</i>					
GPT-4.1 Mini	163.78 \pm 98.56	86.79 \pm 60.53	186.19 \pm 107.64	111.20 \pm 77.52	136.99
+CoT	282.98 \pm 100.23	206.23 \pm 69.57	302.09 \pm 99.34	216.46 \pm 80.95	251.94
<i>Open-weight Large Language Models</i>					
Qwen2.5 3B Instruct	42.18 \pm 36.45	24.36 \pm 20.55	46.25 \pm 38.97	33.09 \pm 30.90	36.47
+CoT	180.81 \pm 329.07	157.29 \pm 350.21	176.33 \pm 188.08	153.23 \pm 67.16	166.92
Qwen2.5 7B Instruct	30.90 \pm 18.33	19.26 \pm 13.26	33.21 \pm 21.03	26.23 \pm 18.85	27.40
+CoT	4705.88 \pm 3976.94	3723.11 \pm 4008.96	4801.62 \pm 3962.49	4504.85 \pm 4015.31	4433.86
Qwen2.5 14B Instruct	37.05 \pm 20.16	26.90 \pm 16.62	42.33 \pm 25.19	31.74 \pm 21.19	34.51
+CoT	167.78 \pm 54.69	142.31 \pm 248.58	178.23 \pm 58.97	162.10 \pm 422.30	162.61
Qwen3 4B Instruct 2507	61.19 \pm 48.45	98.58 \pm 676.35	69.45 \pm 93.76	47.98 \pm 39.37	69.30
+CoT	490.88 \pm 778.15	431.86 \pm 1009.89	500.40 \pm 647.48	409.64 \pm 796.57	458.19
Gemma 3 4B IT	60.97 \pm 88.76	56.17 \pm 71.18	57.65 \pm 68.12	70.92 \pm 310.54	61.43
+CoT	272.78 \pm 568.13	228.08 \pm 566.71	291.60 \pm 637.28	210.28 \pm 336.13	250.69
<i>Open-weight Medical LLMs</i>					
MedGemma 4B IT	885.26 \pm 2464.30	1032.49 \pm 2662.78	966.32 \pm 2566.97	921.43 \pm 2523.40	951.38
+CoT	8179.20 \pm 310.00	8177.36 \pm 334.93	8180.46 \pm 297.15	8181.04 \pm 293.15	8179.52
MedGemma 27B IT	75.15 \pm 363.27	33.21 \pm 249.06	66.23 \pm 202.91	33.26 \pm 42.59	51.96
+CoT	1011.49 \pm 842.55	945.87 \pm 1202.11	1237.82 \pm 1249.42	866.86 \pm 904.18	1015.51
<i>Open-weight Reasoning Models</i>					
OpenThinker3 7B	4915.88 \pm 3022.94	4411.53 \pm 3028.07	5495.49 \pm 2829.85	4154.27 \pm 3049.39	4744.29
+CoT	7842.04 \pm 1404.18	7561.84 \pm 1881.72	7737.29 \pm 1577.02	7377.80 \pm 2054.35	7629.74
<i>Open-weight Medical Reasoning Models</i>					
HuatuoGPT o1 7B	55.94 \pm 330.24	34.55 \pm 176.57	61.56 \pm 358.68	36.49 \pm 31.11	47.14
+CoT	395.63 \pm 275.05	361.73 \pm 326.84	404.67 \pm 150.04	336.29 \pm 335.71	374.58
m1 7B 23K	1629.10 \pm 2926.38	1564.04 \pm 2882.66	1733.96 \pm 2945.14	1702.78 \pm 2966.04	1657.47
+CoT	8146.77 \pm 548.03	8149.01 \pm 552.57	8166.48 \pm 418.94	8133.05 \pm 648.74	8148.83
AlphaMed 7B Instruct RL	20.58 \pm 28.28	17.68 \pm 175.47	20.46 \pm 32.36	17.26 \pm 17.24	19.00
+CoT	1340.47 \pm 2803.22	852.62 \pm 2297.19	1033.94 \pm 2461.34	979.52 \pm 2451.48	1051.64

2527

2528

2529

2530

2531

2532

2533

2534

2535

2536

2537

2538
 2539
 2540
 2541
 2542
 2543
 2544
 2545
 2546
 2547

2548 Table 21: Metrics related to the ranked **list** answer format from the generated evaluation responses
 2549 across benchmarks from Section 3. **CP** denotes the average rank position of the correct item within
 2550 the generated list, **LL** represents the average list length across all responses, and **VLL** corresponds
 2551 to the average valid list length, computed only over non-empty outputs.

	MedQA			MedMCQA			MXQA			MLUP-H		
	LL	VLL	CP	LL	VLL	CP	LL	VLL	CP	LL	VLL	CP
<i>Proprietary Reasoning Models</i>												
Gemini 2.5 Flash Lite	2.84	3.00	1.27	2.70	2.73	1.46	2.96	3.07	1.59	2.59	2.61	1.22
+CoT	0.75	2.05	1.10	1.16	1.76	1.17	0.60	2.11	1.32	0.96	1.74	1.09
Gemini 2.5 Flash	3.26	3.31	1.34	2.42	2.42	1.44	3.39	3.42	1.53	2.86	2.87	1.32
+CoT	1.08	2.92	1.30	1.33	2.28	1.26	0.92	2.93	1.34	1.30	2.53	1.28
Gemini 2.5 Pro	3.66	3.71	1.27	3.02	3.02	1.54	3.58	3.62	1.48	3.28	3.30	1.30
+CoT	3.56	3.69	1.28	2.92	3.09	1.56	3.51	3.69	1.47	3.20	3.37	1.34
<i>Proprietary Large Language Models</i>												
GPT-4.1 Mini	3.44	3.59	1.21	2.90	2.90	1.48	3.40	3.46	1.50	3.07	3.09	1.24
+CoT	3.99	4.00	1.22	3.39	3.39	1.51	3.88	3.88	1.49	3.59	3.59	1.22
<i>Open-weight Large Language Models</i>												
Qwen2.5 3B Instruct	3.56	3.56	1.67	2.42	2.42	1.57	3.53	3.53	1.95	2.86	2.86	1.56
+CoT	1.79	1.83	1.24	1.61	1.61	1.25	1.90	1.93	1.35	1.83	1.84	1.23
Qwen2.5 7B Instruct	2.73	2.73	1.43	1.84	1.84	1.34	2.68	2.68	1.61	2.29	2.29	1.41
+CoT	195.49	196.93	1.65	153.75	154.89	2.04	210.52	212.15	2.25	176.98	178.92	1.70
Qwen2.5 14B Instruct	3.23	3.25	1.43	2.76	2.76	1.52	3.28	3.29	1.70	2.89	2.89	1.29
+CoT	2.57	2.57	1.27	2.07	2.07	1.28	2.57	2.57	1.49	2.24	2.24	1.17
Qwen3 4B Instruct 2507	4.42	4.42	1.68	3.40	3.40	1.75	4.38	4.38	1.90	3.64	3.64	1.48
+CoT	3.75	3.95	1.41	2.69	2.93	1.53	3.76	3.96	1.74	3.10	3.41	1.32
Gemma 3 4B IT	4.40	4.58	1.78	4.00	4.02	1.94	4.29	4.40	1.94	5.66	5.72	1.64
+CoT	3.60	4.22	1.58	2.94	3.63	1.75	3.79	4.64	1.78	3.37	4.03	1.57
<i>Open-weight Medical LLMs</i>												
MedGemma 4B IT	90.18	90.33	2.12	104.13	104.13	2.33	87.22	87.31	2.43	98.49	98.49	1.72
+CoT	491.91	546.91	2.44	331.35	387.53	3.30	429.49	515.78	3.54	423.99	480.82	3.26
MedGemma 27B IT	3.66	3.82	1.37	2.99	2.99	1.48	3.29	3.36	1.72	2.86	2.89	1.28
+CoT	3.95	4.01	1.31	3.14	3.21	1.53	4.70	4.83	1.61	3.44	3.49	1.25
<i>Open-weight Reasoning Models</i>												
OpenThinker3 7B	2.25	3.45	1.36	1.99	2.73	1.38	1.84	3.04	1.60	2.17	3.01	1.37
+CoT	0.07	1.84	1.00	0.13	2.30	1.83	0.57	13.09	1.00	0.16	1.80	1.33
<i>Open-weight Medical Reasoning Models</i>												
HuatuoGPT o1 7B	3.14	5.36	1.58	3.21	3.76	1.84	2.46	4.81	1.89	2.41	3.64	1.50
+CoT	0.04	2.61	1.16	0.03	1.99	1.54	0.03	3.12	2.00	0.02	2.07	1.12
m1 7B 23K	14.44	14.44	1.85	9.82	9.82	2.06	18.23	18.23	2.36	11.06	11.06	1.77
+CoT	13.00	17.14	1.45	14.85	18.01	1.83	16.22	22.58	1.89	14.42	18.39	1.37
AlphaMed 7B Instruct RL	2.00	2.68	1.92	1.58	2.35	1.72	2.03	2.73	2.08	2.11	2.61	1.58
+CoT	61.01	71.04	1.68	35.16	40.67	1.84	50.63	60.62	2.42	44.65	50.64	1.60

2582
 2583
 2584
 2585
 2586
 2587
 2588
 2589
 2590
 2591

2592 Table 22: Results table for SFT experiments evaluated on the benchmarks with **MCQ** answer format
 2593 from Section 4.1. MXQA refers to MedXpertQA (text), and MLUP-H refers to MMLU Pro (Health).

	MedQA	MedMCQA	MXQA	MLUP-H	Average
SFT-MCQ	66.46	18.44	14.20	59.29	39.60
+CoT	70.23	19.09	15.71	61.12	41.54
SFT-QA	67.32	15.59	13.63	54.16	37.67
+CoT	66.06	16.08	13.18	55.87	37.80
SFT-List	26.24	0.31	5.67	10.51	10.68
+CoT	29.85	0.46	6.61	13.57	12.62

2601 Table 23: Results table for SFT experiments evaluated on the benchmarks with **QA** answer format
 2602 from Section 4.1. MXQA refers to MedXpertQA (text), and MLUP-H refers to MMLU Pro (Health).

	MedQA		MedMCQA		MXQA		MLUP-H		Average	
	Acc _{QA}	Acc _{LLM} ^{QA}	Acc _{QA} ^{LLM}							
SFT-QA	18.25	52.72	7.75	42.48	5.27	40.41	13.72	50.68	11.25	46.57
+CoT	18.09	53.28	8.35	42.11	4.89	40.17	13.72	52.58	11.26	47.04
SFT-MCQ	17.19	53.69	8.26	43.53	5.42	41.56	13.45	53.40	11.08	48.04
+CoT	19.14	56.85	8.62	43.99	4.84	38.59	14.27	50.82	11.72	47.56
SFT-List	0.32	0.97	0.05	0.64	0.19	1.63	0.41	1.36	0.24	1.15
+CoT	0.41	1.95	0.23	0.78	0.24	1.87	0.82	2.58	0.42	1.80

2612
 2613 per benchmark. Table 21 shows metrics related to list responses, e.g., list length and position of a
 2614 correct item.

2616 N.2 FINE-TUNING

2618 N.2.1 SFT

2620 Tables 22 to 24 present the non-aggregated versions of the aggregated results shown in the main
 2621 body of this study (Table 1 in Section 4.1). While, Tables 26 to 28 show average response length
 2622 per benchmark. Table 29 shows metrics related to list responses, e.g., list length and position of a
 2623 correct item.

2625 N.2.2 RFT

2626 Tables 30 to 32 present the non-aggregated versions of the aggregated results shown in the main
 2627 body of this study (Table 1 in Section 4.2). While, Tables 33 to 35 show average response length
 2628 per benchmark. Table 36 shows metrics related to list responses, e.g., list length and position of a
 2629 correct item.

2632 O QUALITATIVE EXAMPLES

2634 Figure 28 shows an example in which HuatuoGPT-o1 fails to follow the instruction to place the
 2635 final answer inside a boxed environment for the question: “A junior orthopaedic surgery resident is
 2636 completing a carpal tunnel repair with the department chairman as the attending physician. During
 2637 the case, the resident inadvertently cuts a flexor tendon. The tendon is repaired without complication.”

2638 Table 24: Results table for SFT experiments evaluated on the benchmarks with a ranked **list** answer
 2639 format from Section 4.1. MXQA refers to MedXpertQA (text), and MLUP-H refers to MMLU Pro
 2640 (Health).

	MedQA				MedMCQA				MXQA				MLUP-H				Average			
	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}	Acc _{List}	Acc _{List} ^{LLM}	MRR _{List}	MRR _{List} ^{LLM}
SFT-List	24.90	61.64	22.40	54.89	11.33	36.42	9.76	30.32	8.92	37.73	6.96	29.40	15.90	60.60	14.83	54.30	15.26	49.10	13.49	42.23
+CoT	24.01	60.50	21.63	53.91	11.10	37.75	9.71	31.38	8.96	37.87	7.00	28.86	15.76	60.19	14.79	52.27	14.96	49.08	13.28	41.60
SFT-MCQ	20.84	63.42	15.36	45.91	15.50	51.28	10.18	33.16	9.97	46.26	6.36	28.67	15.62	69.02	12.79	51.00	15.48	57.50	11.17	39.68
+CoT	17.84	47.69	17.43	46.31	8.39	25.37	8.37	24.59	5.90	22.58	13.86	47.15	13.86	46.00	11.50	36.09	11.34	34.87		
SFT-QA	18.41	54.66	15.22	44.62	12.84	46.56	9.22	32.77	7.19	38.30	5.31	27.17	13.32	65.76	11.15	51.53	12.94	51.32	10.22	39.02
+CoT	12.49	31.87	12.17	30.62	6.01	18.39	5.82	17.54	3.64	14.81	3.43	13.36	9.38	33.29	9.24	32.22	7.88	24.59	7.67	23.44

2646
 2647 Table 25: Results table for prompting experiments evaluated on open-ended benchmarks using **List**
 2648 prompting strategies. The highest overall score in each benchmark column is shown in **bold**.

2649 2650 Model	2651 MedQA		2652 MedMCQA		2653 MXQA		2654 MLUP-H		2655 Average	
	2656 Acc ^{LLM} _{List}	2657 MRR ^{LLM} _{List}	2658 Acc ^{LLM} _{List}	2659 MRR ^{LLM} _{List}	2660 Acc ^{LLM} _{List}	2661 MRR ^{LLM} _{List}	2662 Acc ^{LLM} _{List}	2663 MRR ^{LLM} _{List}	2664 Acc ^{LLM} _{List}	2665 MRR ^{LLM} _{List}
SFT-List	61.64	54.89	36.42	30.32	37.73	29.40	60.60	54.30	49.10	42.23
SFT-MCQ	63.42	45.91	51.28	33.16	46.26	28.67	69.02	51.00	57.50	39.68
SFT-QA	54.66	44.62	46.56	32.77	38.30	27.17	65.76	51.53	51.32	39.02

2656
 2657 Table 26: Response length (mean \pm standard deviation) for **MCQ** answer format for the experiments
 2658 from Section 4.1.

2659 2660 Model	2661 MedQA		2662 MedMCQA		2663 MXQA		2664 MLUP-H		2665 Average				
	2666 SFT-MCQ	2667 +CoT	2668 3332.74 \pm 1671.04	2669 1344.50 \pm 1316.94	2670 3918.01 \pm 1859.36	2671 2399.41 \pm 1657.66	2672 2748.66	2673 SFT-QA	2674 +CoT	2675 2933.34 \pm 1710.83	2676 1267.93 \pm 1394.58	2677 3152.43 \pm 1819.12	2678 2105.04 \pm 1657.94
SFT-List	3656.68 \pm 1433.71	1218.50 \pm 1504.40	4169.00 \pm 1664.91	2378.76 \pm 1851.69	2855.74	3525.13 \pm 1581.85	1257.90 \pm 1481.72	4063.66 \pm 1674.81	2256.44 \pm 1626.20	2775.78			

2666
 2667 Table 27: Response length (mean \pm standard deviation) for **QA** answer format for the experiments
 2668 from Section 4.1.

2670 2671 Model	2672 MedQA		2673 MedMCQA		2674 MXQA		2675 MLUP-H		2676 Average									
	2677 SFT-QA	2678 +CoT	2679 1295.77 \pm 991.87	2680 1450.64 \pm 1084.56	2681 1082.37 \pm 1028.77	2682 1195.57 \pm 1168.67	2683 1468.30 \pm 1235.48	2684 1092.76 \pm 923.33	2685 1234.80	2686 SFT-MCQ	2687 +CoT	2688 1448.70 \pm 1168.88	2689 1442.76 \pm 1233.05	2690 1123.19 \pm 978.47	2691 1146.67 \pm 1069.65	2692 1823.66 \pm 1542.80	2693 1281.55 \pm 1125.39	2694 1419.28
SFT-List	1417.71 \pm 1231.00	1019.12 \pm 1068.58	1579.55 \pm 1346.94	1133.24 \pm 1090.30	1287.40	1488.67 \pm 1253.06	1073.88 \pm 1056.35	1648.48 \pm 1418.77	1120.35 \pm 883.86	1332.85								

2678
 2679 Table 28: Response length (mean \pm standard deviation) for a ranked-list answer format for the
 2680 experiments from Section 4.1.

2681 2682 Model	2683 MedQA		2684 MedMCQA		2685 MXQA		2686 MLUP-H		2687 Average									
	2688 SFT-List	2689 +CoT	2690 1365.25 \pm 1074.55	2691 1527.38 \pm 1109.57	2692 1154.27 \pm 960.43	2693 1225.99 \pm 1073.97	2694 1485.45 \pm 1156.37	2695 1704.15 \pm 1294.16	2696 1120.23 \pm 775.11	2697 1281.30	2698 SFT-MCQ	2699 +CoT	2700 1249.83 \pm 3257.88	2701 1801.25 \pm 1373.75	2702 1948.54 \pm 3239.73	2703 2984.07 \pm 3433.67	2704 2320.56 \pm 3310.11	2705 2437.75
SFT-QA	1634.33 \pm 2060.58	1835.73 \pm 2674.40	1798.66 \pm 2182.52	1761.27 \pm 2501.29	1757.50	1629.20 \pm 1134.62	1258.18 \pm 1244.22	1913.98 \pm 1349.85	1325.79 \pm 960.62	1531.79								

2689
 2690 Table 29: Metrics related to the ranked **list** answer format from the generated evaluation responses
 2691 across benchmarks from Section 4.1.

2692 2693 Model	2694 MedQA			2695 MedMCQA			2696 MXQA			2697 MLUP-H		
	2698 LL	2699 VLL	2700 CP	2701 LL	2702 VLL	2703 CP	2704 LL	2705 VLL	2706 CP	2707 LL	2708 VLL	2709 CP
SFT-List	2.72	2.74	1.29	2.11	2.13	1.46	2.82	2.84	1.60	2.37	2.37	1.28
+CoT	2.78	2.80	1.28	2.10	2.13	1.44	2.85	2.88	1.63	2.47	2.49	1.33
SFT-MCQ	144.84	146.14	2.65	135.98	136.04	2.82	151.28	151.79	3.41	132.36	132.90	2.43
+CoT	1.51	1.54	1.07	1.29	1.31	1.07	1.46	1.50	1.16	1.47	1.50	1.07
SFT-QA	8.42	8.75	1.64	15.52	15.70	1.96	13.03	13.66	2.01	17.01	17.27	1.65
+CoT	1.17	2.00	1.09	0.97	1.51	1.13	0.99	1.93	1.28	1.03	1.67	1.08

2700
 2701
 2702
 2703
 2704
 2705
 2706
 2707
 2708
 2709
 2710
 2711
 2712
 2713
 2714
 2715

Table 30: Results table for RFT experiments evaluated on the benchmarks with MCQ answer format from Section 4.2.

	Dataset	Prior Prompt	Reward Function	Prompting	MedQA	MedMCQA	MXQA	MLUP-H	Average
RFT-MCQ (Baseline, No Prompt)	MCQ	No Prompt	MCQ-Acc	MCQ-CoT	66.30	16.99	14.20	58.31	38.95
RFT-MCQ (Baseline, MCQ)	MCQ	MCQ	MCQ-Acc	MCQ	67.48	18.73	14.33	58.68	39.80
RFT-MCQ (MCQ-CoT)	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	66.93	17.45	14.16	58.80	39.34
RFT-QA (Baseline, No Prompt)	QA	No Prompt	QA-Acc	MCQ-CoT	47.05	10.54	12.49	40.95	27.76
RFT-QA (Baseline, QA)	QA	QA	QA-Acc	MCQ	51.06	14.93	10.73	36.19	28.23
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	MCQ	16.03	6.21	8.12	21.52	12.97
RFT-List (List-CoT, Acc)	QA	List-CoT	List-Acc	MCQ-CoT	36.53	7.97	9.55	35.57	22.40
RFT-List (List, MRR)	QA	List	List-MRR	MCQ	12.18	5.79	7.18	16.50	10.41
RFT-List (List-CoT, MRR)	QA	List-CoT	List-MRR	MCQ-CoT	23.96	10.65	7.51	30.81	18.23
RFT-List (List-CoT, Judge-MRR)	QA	List-CoT	LLM-List-MRR	MCQ-CoT	30.09	11.37	5.55	34.96	20.49
RFT-MCQ (MCQ-CoT, 4 Epochs)	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	68.34	18.24	14.37	58.92	39.97
RFT-MCQ (MCQ-CoT, No format)	MCQ	No Prompt	MCQ-Acc-NF	MCQ-CoT	67.56	18.18	13.96	58.56	39.57
RFT-List (List-CoT, Judge-MRR-Simple)	QA	List-CoT	LLM-List-MRR-Simple	MCQ-CoT	24.27	12.15	7.76	31.91	19.02
RFT-List (List-CoT, Judge-MRR-Gemini)	QA	List-CoT	LLM-List-MRR-Gemini	MCQ-CoT	54.12	18.46	9.76	50.12	33.12
RFT-MCQ (Acc) + QA (Acc)	MCQ,QA	MCQ-CoT,QA-CoT	MCQ-Acc,QA-Acc	MCQ-CoT	28.04	0.00	9.88	13.33	12.81
RFT-MCQ (Acc) + List (Acc)	MCQ,QA	MCQ-CoT,List-CoT	MCQ-Acc,List-Acc	MCQ-CoT	68.03	17.84	14.37	58.56	39.70
RFT-MCQ (Acc) + List (MRR)	MCQ,QA	MCQ-CoT,List-CoT	MCQ-Acc,List-MRR	MCQ-CoT	68.74	17.93	13.59	59.78	40.01
Qwen2.5 3B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	54.05	14.96	12.33	43.77	31.28
Qwen2.5 3B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	52.95	25.09	11.43	41.08	32.64
Qwen2.5 3B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	52.24	24.67	10.78	40.95	32.16
Qwen2.5 3B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	50.67	24.70	11.59	41.08	32.01
Qwen3 4B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	75.88	18.81	18.12	68.09	45.22
Qwen3 4B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	75.88	19.80	17.59	67.36	45.16
Qwen3 4B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	73.76	20.07	15.67	65.40	43.73
Qwen3 4B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ	75.18	19.37	16.82	65.89	44.32
OpenThinker3 7B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	56.79	17.11	13.02	48.04	33.74
OpenThinker3 7B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	36.64	21.66	11.51	46.21	34.01
OpenThinker3 7B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	36.25	22.57	11.02	47.07	34.23
OpenThinker3 7B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	37.50	21.76	11.35	47.80	34.60
m1 7B 23K: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	76.28	19.95	19.63	64.06	44.98
m1 7B 23K: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	72.51	21.79	16.33	61.61	43.06
m1 7B 23K: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	73.06	21.69	17.96	63.45	44.04
m1 7B 23K: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	75.88	20.60	16.90	60.76	43.54
AlphaMed 7B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	71.88	17.58	17.06	61.61	42.03
AlphaMed 7B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	71.25	17.06	15.71	62.96	41.74
AlphaMed 7B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	68.19	15.50	12.61	57.09	38.35
AlphaMed 7B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	60.57	14.75	14.20	56.23	36.44
RFT-List (List-CoT, MRR, LP=0)	QA	List-CoT	List-MRR-0	MCQ-CoT	17.12	8.11	5.10	21.39	12.93
RFT-List (List-CoT, MRR, LP=0.1)	QA	List-CoT	List-MRR-0.1	MCQ-CoT	65.04	20.00	12.94	52.93	37.73
RFT-List (List-CoT, MRR, LP=0.3)	QA	List-CoT	List-MRR-0.3	MCQ-CoT	55.22	18.07	10.53	50.24	33.52
RFT-List (List-CoT, MRR, LP=0.5)	QA	List-CoT	List-MRR-0.5	MCQ-CoT	63.00	20.70	12.65	54.77	37.78
RFT-List (List-CoT, MRR, LP=0.7)	QA	List-CoT	List-MRR-0.7	MCQ-CoT	56.48	18.78	12.24	51.34	34.71
RFT-List (List-CoT, MRR, LP=0.9)	QA	List-CoT	List-MRR-0.9	MCQ-CoT	60.88	19.01	11.51	50.49	35.47
AlphaMed 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	MCQ-CoT	23.88	11.72	7.67	44.01	21.82
m1 7B 23K: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	MCQ-CoT	74.78	21.50	17.27	61.25	43.70
OpenThinker3 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	MCQ-CoT	56.64	21.76	11.02	48.04	34.36
Qwen2.5 3B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	MCQ-CoT	49.57	21.38	11.88	41.81	31.16

2741
 2742
 2743
 2744
 2745
 2746
 2747
 2748
 2749
 2750
 2751
 2752
 2753

Table 31: Results table for RFT experiments evaluated on the benchmarks with QA answer format from Section 4.2.

Dataset	Prior Prompt	Reward Function	Prompting	MedQA		MedMCQA		MQXA		MLUP-H		Average		
				Acc _{QA}	Acc _{LM}									
RFT-QA (Baseline, No Prompt)	QA	No Prompt	QA-Acc	5.84	28.71	3.26	31.74	1.53	24.02	5.71	32.20	4.08	29.17	
	QA	QA	QA-Acc	0.65	20.60	1.38	31.93	0.29	19.42	1.49	27.15	0.95	24.78	
	QA	QA	QA-Acc	1.05	21.74	1.28	30.87	0.14	16.78	2.31	26.49	1.20	23.97	
	QA	QA-CoT	QA-Acc	1.30	22.38	1.33	31.51	0.29	19.13	1.22	27.85	1.03	25.22	
RFT-MCQA (Baseline, No Prompt)	MCQ	No Prompt	MCQ-ACC	14.27	51.50	7.39	44.40	5.08	38.59	12.91	50.00	9.91	46.12	
	MCQ	MCQ	MCQ-ACC	12.98	47.69	6.93	44.27	3.74	36.53	13.18	49.86	9.21	44.59	
	MCQ	MCQ	MCQ-ACC	13.38	47.20	7.48	44.77	4.27	37.49	11.82	47.01	9.24	44.12	
	MCQ	MCQ-CoT	MCQ-ACC	14.76	50.36	7.20	45.32	4.41	37.06	13.45	52.58	9.96	46.33	
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	13.54	44.93	6.79	43.67	4.36	37.63	13.04	51.63	9.43	44.46	
	QA	List	List-Acc	12.41	41.12	5.50	36.33	3.79	33.41	10.73	41.98	8.11	38.21	
	QA	List-CoT	List-Acc	6.24	20.03	1.97	17.48	1.58	14.48	7.34	24.05	4.28	19.01	
	QA	List	List-MRR	15.49	50.69	7.48	44.13	4.65	38.11	13.86	51.90	10.37	46.21	
RFT-List (List, MRR)	QA	List	List-MRR	17.52	50.53	7.89	44.27	4.51	37.15	13.45	51.49	10.84	45.86	
	QA	List-CoT	List-MRR	5.68	20.60	3.21	27.75	1.92	16.68	6.93	27.58	4.41	21.90	
	QA	List	LLM-List-MRR	9.73	30.49	4.77	30.28	2.78	25.80	8.70	35.05	6.50	30.36	
	QA	List-CoT	LLM-List-MRR	14.76	50.36	7.20	45.32	4.41	37.06	13.45	52.58	9.96	46.33	
RFT-MCQ (MCQ-CoT, 4 Epochs)	MCQ	MCQ-CoT	MCQ-ACC	15.09	48.66	7.29	43.99	4.65	37.68	12.06	50.00	9.78	45.07	
	MCQ	MCQ-CoT	MCQ-ACC	13.87	48.74	7.34	44.91	4.36	39.19	12.36	52.17	9.48	46.24	
	MCQ	No Prompt	MCQ-ACC-NF	QA-CoT	13.87	48.74	7.34	44.91	4.36	39.19	12.36	52.17	9.48	46.24
	MCQ	No Prompt	MCQ-ACC-NF	QA-CoT	13.87	48.74	7.34	44.91	4.36	39.19	12.36	52.17	9.48	46.24
RFT-List (List-CoT, Judge-MRR-Simple)	QA	List-CoT	LLM-List-MRR-Simple	QA-CoT	4.70	26.85	4.13	28.35	1.87	21.96	6.79	31.93	4.37	27.27
	QA	List-CoT	LLM-List-MRR-Gemini	QA-CoT	14.19	46.96	7.39	42.20	4.22	34.71	11.82	48.78	9.40	43.16
	QA	List-CoT	LLM-List-MRR-Gemini	QA-CoT	14.19	46.96	7.39	42.20	4.22	34.71	11.82	48.78	9.40	43.16
	QA	List-CoT	LLM-List-MRR-Gemini	QA-CoT	14.19	46.96	7.39	42.20	4.22	34.71	11.82	48.78	9.40	43.16
RFT-MCQ (QA) + Acc (QA)	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	4.70	29.85	3.30	37.61	0.96	23.54	4.62	36.82	3.40	31.96	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
RFT-MCQ (QA) + Acc (List)	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
RFT-MCQ (QA) + Acc (List, MRR)	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
	MCQ-QA	MCQ-QA-CoT	MCQ-ACC	16.79	52.47	8.53	44.04	5.80	40.36	13.86	50.68	11.24	46.89	
RFT-List (QA, 3BR, RFT-MCQ)	QA	List	MCQ-CoT	8.92	35.36	3.81	33.17	9.24	38.45	8.62	36.45	9.78	34.45	
	QA	List	MCQ-CoT	1.46	20.03	3.07	33.47	0.19	32.91	7.25	30.29	0.87	27.68	
	QA	List	MCQ-CoT	11.02	40.96	5.73	40.31	2.87	38.51	9.89	40.08	7.25	30.29	
	QA	List	MCQ-CoT	11.11	40.71	6.38	41.19	3.21	36.53	11.01	42.39	7.29	40.20	
RFT-List (QA, 3BR, RFT-MCQ)	QA	List	MCQ-CoT	17.71	51.09	8.53	45.05	5.18	39.55	13.18	48.51	11.00	46.05	
	QA	List	MCQ-CoT	12.90	48.99	7.48	44.91	3.69	35.04	10.19	50.95	8.57	44.97	
	QA	List	MCQ-CoT	18.57	53.28	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
	QA	List	MCQ-CoT	18.57	53.45	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	8.92	35.36	3.81	33.17	9.24	38.45	8.62	36.45	9.78	34.45	
	QA	List	MCQ-CoT	1.46	20.03	3.07	33.47	0.19	32.91	7.25	30.29	0.87	27.68	
	QA	List	MCQ-CoT	11.02	40.96	5.73	40.31	2.87	38.51	9.89	40.08	7.25	30.29	
	QA	List	MCQ-CoT	11.11	40.71	6.38	41.19	3.21	36.53	11.01	42.39	7.29	40.20	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	17.71	51.09	8.53	45.05	5.18	39.55	13.18	48.51	11.00	46.05	
	QA	List	MCQ-CoT	12.90	48.99	7.48	44.91	3.69	35.04	10.19	50.95	8.57	44.97	
	QA	List	MCQ-CoT	18.57	53.28	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
	QA	List	MCQ-CoT	18.57	53.45	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	8.92	35.36	3.81	33.17	9.24	38.45	8.62	36.45	9.78	34.45	
	QA	List	MCQ-CoT	1.46	20.03	3.07	33.47	0.19	32.91	7.25	30.29	0.87	27.68	
	QA	List	MCQ-CoT	11.02	40.96	5.73	40.31	2.87	38.51	9.89	40.08	7.25	30.29	
	QA	List	MCQ-CoT	11.11	40.71	6.38	41.19	3.21	36.53	11.01	42.39	7.29	40.20	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	17.71	51.09	8.53	45.05	5.18	39.55	13.18	48.51	11.00	46.05	
	QA	List	MCQ-CoT	12.90	48.99	7.48	44.91	3.69	35.04	10.19	50.95	8.57	44.97	
	QA	List	MCQ-CoT	18.57	53.28	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
	QA	List	MCQ-CoT	18.57	53.45	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	8.92	35.36	3.81	33.17	9.24	38.45	8.62	36.45	9.78	34.45	
	QA	List	MCQ-CoT	1.46	20.03	3.07	33.47	0.19	32.91	7.25	30.29	0.87	27.68	
	QA	List	MCQ-CoT	11.02	40.96	5.73	40.31	2.87	38.51	9.89	40.08	7.25	30.29	
	QA	List	MCQ-CoT	11.11	40.71	6.38	41.19	3.21	36.53	11.01	42.39	7.29	40.20	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	17.71	51.09	8.53	45.05	5.18	39.55	13.18	48.51	11.00	46.05	
	QA	List	MCQ-CoT	12.90	48.99	7.48	44.91	3.69	35.04	10.19	50.95	8.57	44.97	
	QA	List	MCQ-CoT	18.57	53.28	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
	QA	List	MCQ-CoT	18.57	53.45	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	8.92	35.36	3.81	33.17	9.24	38.45	8.62	36.45	9.78	34.45	
	QA	List	MCQ-CoT	1.46	20.03	3.07	33.47	0.19	32.91	7.25	30.29	0.87	27.68	
	QA	List	MCQ-CoT	11.02	40.96	5.73	40.31	2.87	38.51	9.89	40.08	7.25	30.29	
	QA	List	MCQ-CoT	11.11	40.71	6.38	41.19	3.21	36.53	11.01	42.39	7.29	40.20	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	17.71	51.09	8.53	45.05	5.18	39.55	13.18	48.51	11.00	46.05	
	QA	List	MCQ-CoT	12.90	48.99	7.48	44.91	3.69	35.04	10.19	50.95	8.57	44.97	
	QA	List	MCQ-CoT	18.57	53.28	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
	QA	List	MCQ-CoT	18.57	53.45	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	8.92	35.36	3.81	33.17	9.24	38.45	8.62	36.45	9.78	34.45	
	QA	List	MCQ-CoT	1.46	20.03	3.07	33.47	0.19	32.91	7.25	30.29	0.87	27.68	
	QA	List	MCQ-CoT	11.02	40.96	5.73	40.31	2.87	38.51	9.89	40.08	7.25	30.29	
	QA	List	MCQ-CoT	11.11	40.71	6.38	41.19	3.21	36.53	11.01	42.39	7.29	40.20	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	17.71	51.09	8.53	45.05	5.18	39.55	13.18	48.51	11.00	46.05	
	QA	List	MCQ-CoT	12.90	48.99	7.48	44.91	3.69	35.04	10.19	50.95	8.57	44.97	
	QA	List	MCQ-CoT	18.57	53.28	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
	QA	List	MCQ-CoT	18.57	53.45	9.04	46.15	5.70	40.70	13.45	53.67	11.69	48.45	
RFT-List (QA, 3BR, RFT-MR)	QA	List	MCQ-CoT	8.92	35.36	3.81	33.17	9.24	38.45	8.6				

Table 32: Results table for RFT experiments evaluated on the benchmarks with a ranked-list answer format from Section 4.2.

Dataset	Prior Prompt	Reward Function	Prompting	MeDQA						MeQQA						MLUPH						Average			
				Acc ₁	Acc ₂	MRR ₁₂	MRR ₁₂₃₄	Acc ₁	Acc ₂	MRR ₁₂	MRR ₁₂₃₄	Acc ₁	Acc ₂	MRR ₁₂	MRR ₁₂₃₄	Acc ₁	Acc ₂	MRR ₁₂	MRR ₁₂₃₄	Acc ₁	Acc ₂	MRR ₁₂	MRR ₁₂₃₄		
RFTList (Baseline, List, Acc)	QA	List-Acc	Lis-Acc	33.98	47.61	20.05	44.03	23.26	62.66	11.91	30.76	17.26	56.18	8.32	27.04	22.15	74.86	15.05	46.69	24.16	67.08	13.83	77.73	40.35	46.06
RFTList (Baseline, List, Acc)	QA	List-Acc	Lis-Acc	33.20	46.21	21.80	49.22	17.20	58.10	11.30	32.70	17.51	56.18	15.05	28.02	18.50	75.46	15.82	50.65	18.46	75.46	15.82	50.65		
RFTList (List, Acc)	QA	List-Acc	Lis-Acc	33.20	46.21	21.80	49.22	17.20	58.10	11.30	32.70	17.51	56.18	15.05	28.02	18.50	75.46	15.82	50.65	18.46	75.46	15.82	50.65		
RFTList (List, MRR)	QA	List-Acc	Lis-Acc	30.98	46.91	22.22	48.05	20.48	59.63	12.80	34.36	14.81	52.59	8.74	29.03	21.88	72.55	17.34	53.24	22.06	63.00	15.77	51.47		
RFTList (List, MRR)	QA	List-Acc	Lis-Acc	30.98	46.91	22.22	48.05	20.48	59.63	12.80	34.36	14.81	52.59	8.74	29.03	21.88	72.55	17.34	53.24	22.06	63.00	15.77	51.47		
RFTList (List, CTR, MRR)	QA	List-Acc	Lis-Acc	30.33	70.88	24.14	58.44	18.30	55.24	12.36	27.74	14.37	48.61	9.02	31.75	21.74	74.79	17.89	58.23	20.96	61.60	15.83	44.89		
RFTList (List, CTR, MRR)	QA	List-Acc	Lis-Acc	30.33	70.88	24.14	58.44	18.30	55.24	12.36	27.74	14.37	48.61	9.02	31.75	21.74	74.79	17.89	58.23	20.96	61.60	15.83	44.89		
RFTList (List, CTR, MRR)	QA	List-Acc	Lis-Acc	23.00	73.72	18.51	56.07	12.61	53.62	9.63	39.62	9.73	49.42	6.88	35.64	14.81	73.23	13.60	43.41	14.36	69.70	12.61	48.68		
RFTList (No Prompt)	MCQ	MCQ-Acc	MCQ-Acc	14.11	45.50	12.20	23.35	7.71	28.21	7.12	26.83	6.23	28.95	4.93	12.12	12.09	52.45	11.36	46.18	10.04	38.78	8.93	32.71		
RFTMCQ (Baseline, MCQ)	MCQ	MCQ-Acc	MCQ-Acc	13.93	45.17	12.29	28.82	7.91	21.67	6.71	21.60	3.96	20.42	3.67	17.86	10.60	44.13	10.33	43.27	8.92	32.46	8.42	30.19		
RFTMCQ (Baseline, MCQ)	MCQ	MCQ-Acc	MCQ-Acc	13.93	45.17	12.29	28.82	7.91	21.67	6.71	21.60	3.96	20.42	3.67	17.86	10.60	44.13	10.33	43.27	8.92	32.46	8.42	30.19		
RFTMCQ (No Prompt)	QA	QA-Acc	QA-Acc	15.47	49.80	12.87	23.41	8.39	34.76	7.44	58.02	8.62	47.22	20.94	9.38	54.89	8.73	47.52	9.75	41.92	8.32	34.75	7.44	34.75	
RFTMCQ (No Prompt)	QA	QA-Acc	QA-Acc	16.14	55.47	12.77	42.68	10.64	43.03	8.22	32.14	6.38	45.84	4.53	62.83	12.50	62.77	10.72	51.56	11.42	49.93	9.06	38.35		
RFTMCQ (No Prompt)	QA	QA-Acc	QA-Acc	0.81	3.16	0.63	2.38	0.50	0.34	2.34	0.52	0.82	0.43	1.92	0.82	5.43	0.82	0.43	0.67	3.59	0.48	2.82			
RFTMCQ (MCQ-Cot, 4 Epochs)	MCQ	MCQ-Acc	MCQ-Acc	15.82	45.69	14.90	41.72	7.61	26.15	7.12	23.70	4.51	22.63	4.15	19.49	11.68	47.96	11.29	43.88	9.90	35.56	9.55	32.20		
RFTMCQ (MCQ-Cot, 4 Epochs)	MCQ	MCQ-Acc	MCQ-Acc	15.82	45.69	14.90	41.72	7.61	26.15	7.12	23.70	4.51	22.63	4.15	19.49	11.68	47.96	11.29	43.88	9.90	35.56	9.55	32.20		
RFTMCQ (MCQ-Cot, 4 Epochs)	QA	QA-Acc	QA-Acc	15.82	45.69	14.90	41.72	7.61	26.15	7.12	23.70	4.51	22.63	4.15	19.49	11.68	47.96	11.29	43.88	9.90	35.56	9.55	32.20		
RFTMCQ (MCQ-Cot, 4 Epochs)	QA	QA-Acc	QA-Acc	15.82	45.69	14.90	41.72	7.61	26.15	7.12	23.70	4.51	22.63	4.15	19.49	11.68	47.96	11.29	43.88	9.90	35.56	9.55	32.20		
RFTMCQ (List, Judge-Ret, Gemini)	QA	List-Acc	Lis-Acc	15.82	46.21	14.06	48.15	7.43	26.62	6.87	24.55	5.22	26.46	5.15	21.35	11.96	50.60	16.09	40.71	9.01	37.91	8.37	34.75		
RFTMCQ (List, Judge-Ret, Gemini)	QA	List-Acc	Lis-Acc	15.82	46.21	14.06	48.15	7.43	26.62	6.87	24.55	5.22	26.46	5.15	21.35	11.96	50.60	16.09	40.71	9.01	37.91	8.37	34.75		
RFTMCQ (List, Judge-Ret, Gemini)	QA	QA-Acc	QA-Acc	15.82	46.21	14.06	48.15	7.43	26.62	6.87	24.55	5.22	26.46	5.15	21.35	11.96	50.60	16.09	40.71	9.01	37.91	8.37	34.75		
RFTMCQ (List, Judge-Ret, Gemini)	QA	QA-Acc	QA-Acc	15.82	46.21	14.06	48.15	7.43	26.62	6.87	24.55	5.22	26.46	5.15	21.35	11.96	50.60	16.09	40.71	9.01	37.91	8.37	34.75		
Ques2.5B: RFT-MCQ	MCQ	MCQ-Cot	MCQ-Cot	11.44	38.04	10.48	33.95	6.56	24.17	5.60	32.66	3.24	3.06	1.82	9.38	44.16	8.33	38.71	3.75	26.01	9.86	28.01			
Ques2.5B: RFT-MCQ	MCQ	MCQ-Cot	MCQ-Cot	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	MCQ	MCQ-Cot	MCQ-Cot	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84	44.78	7.54	33.82			
Ques2.5B: RFT-MCQ	QA	QA-Acc	QA-Acc	14.44	57.07	10.82	40.61	8.09	33.95	6.56	32.66	3.24	3.06	1.82	9.38	44.16	8.33	44.80	9.84						

2808
2809
2810
2811Table 33: Response length (mean \pm standard deviation) for MCQ answer format for the experiments from Section 4.2.

	Dataset	Prior Prompt	Reward Function	Prompting	MedQA	MedMCQA	MXOA	MLUP-H	Average
RFT-MCQ (Baseline, No Prompt)	MCQ	No Prompt	MCQ-Acc	MCQ-CoT	520.06 \pm 660.03	293.64 \pm 281.92	573.46 \pm 682.75	497.52 \pm 559.83	473.42
RFT-MCQ (Baseline, MCQ)	MCQ	MCQ	MCQ-Acc	MCQ	176.62 \pm 90.24	77.53 \pm 64.76	141.82 \pm 182.70	116.84 \pm 107.32	128.20
RFT-MCQ (MCQ-CoT)	MCQ	MCQ	MCQ-Acc	MCQ-CoT	231.71 \pm 90.31	173.14 \pm 128.79	222.21 \pm 126.87	263.59 \pm 143.15	222.66
RFT-QA (Baseline, No Prompt)	QA	No Prompt	QA-Acc	MCQ-CoT	330.54 \pm 670.02	171.21 \pm 329.32	302.67 \pm 403.41	229.17 \pm 107.92	258.40
RFT-QA (Baseline, QA)	QA	QA	QA-Acc	MCQ	178.06 \pm 332.86	66.85 \pm 59.55	155.64 \pm 247.44	117.39 \pm 90.03	129.48
RFT-QA (Baseline, QA)	QA	QA	QA-Acc	MCQ-CoT	277.26 \pm 398.19	140.54 \pm 116.14	282.95 \pm 296.63	212.52 \pm 115.77	228.32
RFT-QA (QA-CoT)	QA	QA-CoT	QA-Acc	MCQ-CoT	323.16 \pm 243.93	192.73 \pm 163.01	382.19 \pm 492.29	285.82 \pm 305.12	295.97
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	MCQ	124.63 \pm 242.31	40.48 \pm 51.09	108.21 \pm 86.27	64.19 \pm 73.56	84.38
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	MCQ-CoT	220.93 \pm 74.88	131.90 \pm 54.69	224.68 \pm 186.61	194.70 \pm 96.21	193.05
RFT-List (List-CoT, Acc)	QA	List-CoT	List-Acc	MCQ-CoT	244.73 \pm 326.16	132.18 \pm 113.00	244.32 \pm 185.84	208.52 \pm 295.76	207.44
RFT-List (List, MRR)	QA	List	List-MRR	MCQ	21.51 \pm 47.27	15.78 \pm 28.55	45.85 \pm 176.77	25.70 \pm 46.81	27.21
RFT-List (List, MRR)	QA	List	List-MRR	MCQ-CoT	228.82 \pm 65.15	159.50 \pm 186.67	221.85 \pm 96.12	238.78 \pm 296.46	212.09
RFT-List (List-CoT, MRR)	QA	List-CoT	List-MRR	MCQ-CoT	187.08 \pm 64.94	122.08 \pm 44.46	208.78 \pm 242.88	176.09 \pm 83.81	173.50
RFT-List (List-CoT, Judge-MRR-Gemini)	QA	List-CoT	LLM-List-MRR-Gemini	MCQ-CoT	282.94 \pm 98.54	140.54 \pm 120.62	278.70 \pm 174.24	243.20 \pm 128.07	238.85
RFT-MCQ (MCQ-CoT, 4 Epochs)	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	328.58 \pm 78.29	202.53 \pm 66.22	435.50 \pm 151.71	379.03 \pm 130.02	335.73
RFT-MCQ (MCQ-CoT, No format)	MCQ	No Prompt	MCQ-Acc-NF	MCQ-CoT	206.18 \pm 232.18	132.61 \pm 152.00	190.21 \pm 78.16	179.31 \pm 89.46	177.08
RFT-List (List-CoT, Judge-MRR)	QA	List-CoT	LLM-List-MRR	MCQ-CoT	195.16 \pm 395.85	114.92 \pm 211.49	199.62 \pm 491.15	160.93 \pm 405.08	167.66
RFT-List (List-CoT, Judge-MRR-Simple)	QA	List-CoT	LLM-List-MRR-Simple	MCQ-CoT	212.04 \pm 324.82	100.06 \pm 53.78	236.39 \pm 299.36	188.67 \pm 97.02	191.19
RFT-MCQ (Acc) + QA (Acc)	MCQ,QA	MCQ-CoT,QA-CoT	MCQ-Acc	MCQ-CoT	223.27 \pm 55.45	149.48 \pm 142.09	292.36 \pm 74.87	263.07 \pm 400.57	232.05
RFT-MCQ (Acc) + List (Acc)	MCQ,QA	MCQ-CoT,List-CoT	MCQ-Acc	MCQ-CoT	266.11 \pm 234.10	191.85 \pm 115.36	241.67 \pm 187.56	284.41 \pm 295.77	245.54
RFT-MCQ (Acc) + List (MRR)	MCQ,QA	MCQ-CoT,List-CoT	MCQ-Acc,List-MRR	MCQ-CoT	194.22 \pm 243.90	191.84 \pm 186.25	217.45 \pm 88.60	215.51 \pm 298.04	209.22
Qwen2.5 3B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	211.86 \pm 230.00	153.12 \pm 151.00	219.50 \pm 179.83	203.22 \pm 289.73	196.93
Qwen2.5 3B: RFT-MCQ	QA	QA-CoT	MCQ-Acc	MCQ-CoT	253.43 \pm 232.29	173.06 \pm 151.36	347.80 \pm 333.33	403.54 \pm 263.96	314.66
Qwen2.5 3B: RFT-MCQ	QA	QA-CoT	MCQ-Acc	MCQ-CoT	270.45 \pm 237.46	182.93 \pm 266.59	359.00 \pm 331.46	314.66	333.86
Qwen2.5 3B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	380.28 \pm 455.47	266.84 \pm 82.48	323.41 \pm 264.62	365.62 \pm 139.18	377.05
Qwen3 4B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	967.36 \pm 728.52	508.20 \pm 322.22	1133.19 \pm 626.74	769.45 \pm 562.13	844.55
Qwen3 4B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	999.76 \pm 948.85	583.07 \pm 749.84	1307.53 \pm 948.42	890.22 \pm 845.93	945.14
Qwen3 4B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	929.42 \pm 824.62	493.04 \pm 586.96	1109.69 \pm 718.85	732.17 \pm 625.36	816.09
Qwen3 4B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	1055.11 \pm 919.84	545.52 \pm 622.42	1235.89 \pm 757.93	824.04 \pm 683.32	915.11
OpenThinker3 7B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	1354.63 \pm 662.76	977.25 \pm 611.76	1577.11 \pm 721.77	1344.25 \pm 667.48	1313.24
OpenThinker3 7B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	1256.72 \pm 535.40	966.07 \pm 515.33	1507.42 \pm 625.86	1349.09 \pm 652.15	1269.83
OpenThinker3 7B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	141.56 \pm 81.55	1073.69 \pm 707.64	799.80 \pm 926.08	161.12 \pm 1080.28	1475.05
OpenThinker3 7B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	141.56 \pm 81.55	1073.69 \pm 707.64	799.80 \pm 926.08	161.12 \pm 1080.28	1475.05
m1 7B 3B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	1375.52 \pm 715.10	980.82 \pm 882.88	1826.03 \pm 838.53	1397.08 \pm 854.26	1394.79
m1 7B 3B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	111.09 \pm 618.71	848.73 \pm 584.92	1489.26 \pm 921.50	1278.22 \pm 1049.86	1182.33
m1 7B 3B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	1424.52 \pm 1174.36	1033.81 \pm 1003.63	1867.50 \pm 1173.63	1583.00 \pm 1479.20	1477.21
m1 7B 3B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	1394.42 \pm 837.17	1178.52 \pm 590.39	1958.63 \pm 1097.68	1635.26 \pm 1257.96	1542.71
AlphaMed 7B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	391.99 \pm 502.29	225.45 \pm 157.44	408.06 \pm 373.36	341.60 \pm 110.20	341.72
AlphaMed 7B: RFT-QA	QA	QA-CoT	QA-Acc	MCQ-CoT	306.82 \pm 114.31	172.14 \pm 194.50	279.55 \pm 126.54	263.30 \pm 129.99	255.45
AlphaMed 7B: RFT-List (Acc)	QA	List-CoT	List-Acc	MCQ-CoT	281.47 \pm 150.50	167.01 \pm 150.20	298.01 \pm 178.70	278.01 \pm 115.69	255.63
AlphaMed 7B: RFT-List (MRR)	QA	List-CoT	List-MRR	MCQ-CoT	297.20 \pm 83.25	184.36 \pm 83.76	332.02 \pm 200.01	297.19 \pm 112.16	276.62
RFT-List (List-CoT, MRR, LP=0)	QA	List-CoT	List-MRR	MCQ-CoT	190.76 \pm 231.12	127.77 \pm 182.62	193.30 \pm 72.01	168.04 \pm 77.98	169.97
RFT-List (List-CoT, MRR, LP=0.1)	QA	List-CoT	List-MRR	MCQ-CoT	204.70 \pm 235.12	126.41 \pm 136.36	199.30 \pm 64.44	181.40 \pm 96.79	175.78
RFT-List (List-CoT, MRR, LP=0.3)	QA	List-CoT	List-MRR	MCQ-CoT	191.22 \pm 185.04	111.21 \pm 185.04	190.08 \pm 376.03	145.12 \pm 269.36	153.86
RFT-List (List-CoT, MRR, LP=0.5)	QA	List-CoT	List-MRR	MCQ-CoT	190.67 \pm 76.47	119.39 \pm 48.68	184.64 \pm 81.90	145.40 \pm 93.76	156.32
RFT-List (List-CoT, MRR, LP=0.7)	QA	List-CoT	List-MRR	MCQ-CoT	187.22 \pm 238.88	122.79 \pm 155.43	197.52 \pm 245.17	135.26 \pm 103.58	160.67
RFT-List (List-CoT, MRR, LP=0.9)	QA	List-CoT	List-MRR	MCQ-CoT	104.93 \pm 70.75	23.91 \pm 86.67	63.70 \pm 65.73	46.31 \pm 290.60	59.71
AlphaMed 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc	MCQ-CoT	244.91 \pm 75.56	148.10 \pm 62.38	233.57 \pm 105.36	222.71 \pm 99.02	212.32
m1 7B 3K: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc	MCQ-CoT	204.78 \pm 235.12	126.41 \pm 136.36	199.30 \pm 64.44	181.40 \pm 96.79	175.78
OpenThinker3 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc	MCQ-CoT	192.88 \pm 238.02	126.41 \pm 136.36	199.30 \pm 64.44	181.40 \pm 96.79	175.78
Qwen2.5 3B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc	MCQ-CoT	1298.88 \pm 662.69	953.24 \pm 838.02	1824.68 \pm 118.28	1402.28 \pm 103.36	1369.64
OpenThinker3 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc	MCQ-CoT	1320.86 \pm 668.88	950.61 \pm 564.96	1536.51 \pm 689.73	1334.80 \pm 758.55	1287.50
Qwen2.5 3B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc	MCQ-CoT	222.24 \pm 234.46	159.13 \pm 120.15	251.06 \pm 261.46	285.32 \pm 133.61	229.44
RFT-MCQ (MCQ-CoT, 4 Epochs)	MCQ	MCQ-CoT	MCQ-Acc	MCQ-CoT	229.24 \pm 91.13	168.53 \pm 92.16	243.08 \pm 265.24	199.44 \pm 99.84	210.07
RFT-MCQ (MCQ-CoT, No format)	MCQ	No Prompt	MCQ-Acc-NF	MCQ-CoT	169.14 \pm 54.71	134.18 \pm 49.48	179.33 \pm 61.03	149.91 \pm 61.48	158.14
RFT-List (List-CoT, Judge-MRR-Gemini)	QA	List-CoT	LLM-List-MRR-Gemini	MCQ-CoT	230.34 \pm 340.88	154.86 \pm 61.19	232.24 \pm 87.67	182.32 \pm 77.84	199.94
RFT-List (List-CoT, Judge-MRR-Simple)	QA	List-CoT	LLM-List-MRR-Simple	MCQ-CoT	174.30 \pm 236.07	132.00 \pm 249.02	193.59 \pm 356.67	150.93 \pm 301.78	162.70
RFT-MCQ (Acc) + QA (Acc)	MCQ,QA	MCQ-CoT,QA-CoT	MCQ-Acc	QA-CoT	234.61 \pm 327.94	161.52 \pm 252.33	230.94 \pm 68.45	194.29 \pm 67.50	208.24
RFT-MCQ (Acc) + List (Acc)	MCQ,QA	MCQ-CoT,List-CoT	MCQ-Acc,List-MRR	QA-CoT	209.97 \pm 135.21	181.50 \pm 184.99	212.16 \pm 66.68	197.59 \pm 302.26	200.31
RFT-MCQ (Acc) + List (MRR)	MCQ,QA	MCQ-CoT,List-CoT	MCQ-Acc,List-MRR	QA-CoT	182.23 \pm 236.95	146.83 \pm 253.90	230.31 \pm 746.17	151.78 \pm 424.00	192.16
RFT-MCQ (Acc) + QA (Acc)	MCQ,QA	MCQ-CoT,QA-CoT	MCQ-Acc	QA-CoT	201.90 \pm 232.77	162.63 \pm 42.00	213.34 \pm 307.08	178.58 \pm 50.93	189.11
Qwen2.5 3B: RFT-MCQ	QA	QA-CoT	QA-Acc	QA-CoT	321.63 \pm 251.35	267.13 \pm 237.28	325.50 \pm 307.87	284.53 \pm 301.63	300.46
Qwen2.5 3B: RFT-List (Acc)	QA	List-CoT	List-Acc	QA-CoT	230.47 \pm 243.00	200.84 \pm 254.94	233.08 \pm 201.22	199.69 \pm 80.62	216.02
Qwen2.5 3B: RFT-List (MRR)	QA	List-CoT	List-MRR	QA-CoT	230.74 \pm 96.50	223.32 \pm 87.51	211.82 \pm 99.87	101.04 \pm 313.73	99.98
RFT-List (List-CoT, MRR)	QA	List	List-MRR	QA-CoT	195.21 \pm 231.28	144.55 \pm 63.33	208.19 \pm 311.32	159.03 \pm 67.04	176.64
RFT-List (List-CoT, MRR, LP=0.7)	QA	List-CoT	List-MRR	QA-CoT	172.84 \pm 328.54	145.95 \pm 388.67	165.41 \pm 57.77	139.83 \pm 52.68	156.01
RFT-List (List-CoT, Judge-MRR)	QA	List-CoT	LLM-List-MRR	QA-CoT	242.11 \pm 847				

2862

Table 35: Response length (mean \pm standard deviation) for a ranked-list answer format for the experiments from Section 4.2.

2863

2864

2865

	Dataset	Prior Prompt	Reward Function	Prompting	MedQA	MedMCQA	MXOA	MLUP-H	Average
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	List	7977.79 \pm 1277.05	7717.38 \pm 1891.88	7958.86 \pm 1326.88	7800.39 \pm 1707.57	7863.60
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	List-CoT	3379.40 \pm 3956.56	2825.40 \pm 3826.98	3380.38 \pm 3959.44	3149.64 \pm 3919.37	3183.70
RFT-List (List-CoT, Acc)	QA	List	List-CoT	List-CoT	183.77 \pm 233.69	147.13 \pm 246.98	181.84 \pm 182.46	147.06 \pm 45.10	164.95
RFT-List (List, MRR)	QA	List	List-MRR	List	80.58 \pm 232.6	75.62 \pm 303.56	73.10 \pm 31.02	85.18 \pm 368.82	78.62
RFT-List (List, MRR)	QA	List	List-MRR	List-CoT	123.99 \pm 35.36	101.17 \pm 176.37	130.06 \pm 252.26	115.37 \pm 300.03	117.64
RFT-List (List-CoT, MRR)	QA	List	List-MRR	List-CoT	409.30 \pm 1380.15	265.23 \pm 1059.86	342.08 \pm 1194.10	255.61 \pm 980.80	318.05
RFT-List (List-CoT, Judge-MRR)	QA	List	List-MRR	List-CoT	150.56 \pm 327.42	108.58 \pm 177.57	143.49 \pm 49.30	126.61 \pm 300.57	132.31
RFT-MCQ (Baseline, No Prompt)	MCQ	No Prompt	MCQ-Acc	List-CoT	162.24 \pm 65.98	123.02 \pm 348.54	165.23 \pm 67.44	135.25 \pm 302.00	146.44
RFT-MCQ (Baseline, MCQ)	MCQ	MCQ	MCQ-Acc	List	26.75 \pm 15.78	17.96 \pm 11.82	28.86 \pm 19.72	22.13 \pm 14.45	23.92
RFT-MCQ (Baseline, MCQ)	MCQ	MCQ	MCQ-Acc	List-CoT	122.79 \pm 53.89	80.84 \pm 40.83	130.25 \pm 64.71	96.81 \pm 52.09	107.67
RFT-MCQ (MCQ-CoT)	MCQ	MCQ	MCQ-Acc	List-CoT	34.24 \pm 35.95	29.84 \pm 247.83	37.91 \pm 180.96	27.93 \pm 24.40	32.4
RFT-QA (Baseline, No Prompt)	QA	No Prompt	QA-Acc	List-CoT	180.77 \pm 332.36	129.07 \pm 250.78	169.92 \pm 261.72	140.56 \pm 304.27	155.08
RFT-QA (Baseline, QA)	QA	QA	QA-Acc	List	48.89 \pm 23.26	31.86 \pm 15.97	49.08 \pm 24.28	38.58 \pm 21.30	42.10
RFT-QA (Baseline, QA)	QA	QA	QA-Acc	List-CoT	160.03 \pm 236.0	104.87 \pm 41.05	149.65 \pm 57.38	117.44 \pm 53.06	133.00
RFT-QA (QA-CoT)	QA	QA	QA-Acc	List-CoT	61.37 \pm 239.85	29.37 \pm 27.61	52.42 \pm 51.27	40.43 \pm 44.25	45.90
RFT-MCQ (MCQ-CoT, 4 Epochs)	MCQ	MCQ-CoT	MCQ-Acc	List-CoT	201.07 \pm 87.92	134.61 \pm 187.36	207.87 \pm 197.89	164.63 \pm 90.19	177.04
RFT-MCQ (MCQ-CoT, No format)	MCQ	No Prompt	MCQ-Acc-NF	List-CoT	138.93 \pm 53.27	101.75 \pm 41.95	144.04 \pm 57.74	127.44 \pm 302.30	128.04
RFT-List (List-CoT, Judge-MRR-Simple)	QA	List-CoT	LLM-List-MRR-Simple	List-CoT	156.06 \pm 43.61	114.27 \pm 37.69	159.25 \pm 46.14	126.77 \pm 42.46	139.09
RFT-List (List-CoT, Judge-MRR-Gemini)	QA	List-CoT	LLM-List-MRR-Gemini	List-CoT	160.31 \pm 287.29	113.02 \pm 35.09	157.10 \pm 42.85	132.22 \pm 43.51	140.66
RFT-MCQ (Acc) + QA (Acc)	MCQ, QA	MCQ-CoT, QA-CoT	MCQ-Acc, QA-Acc	List-CoT	137.23 \pm 242.17	103.34 \pm 252.99	121.84 \pm 194.14	101.34 \pm 70.48	115.91
RFT-MCQ (Acc) + List (Acc)	MCQ, QA	MCQ-CoT, QA-CoT	MCQ-Acc, QA-Acc	List-CoT	206.07 \pm 512.63	154.43 \pm 243.94	224.03 \pm 632.92	186.76 \pm 593.83	192.83
RFT-MCQ (Acc) + List (MRR)	MCQ, QA	MCQ-CoT, QA-CoT	MCQ-Acc, List-MRR	List-CoT	143.37 \pm 38.17	126.13 \pm 176.64	144.46 \pm 38.48	123.71 \pm 37.68	132.20
Qwen2.5 3B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	List-CoT	192.78 \pm 63.89	158.83 \pm 181.50	198.30 \pm 62.74	165.53 \pm 67.83	178.86
Qwen2.5 3B: RFT-MCQ	QA	QA-CoT	QA-Acc	List-CoT	31.29 \pm 39.64	34.67 \pm 35.35	35.55 \pm 34.59	30.70 \pm 58.87	32.4
Qwen2.5 3B: RFT-List (Acc)	QA	QA-CoT	QA-Acc	List	70.94 \pm 207.20	70.21 \pm 163.84	70.91 \pm 214.74	77.97 \pm 171.17	80.80
Qwen2.5 3B: RFT-List (MRR)	QA	QA-CoT	QA-Acc	List-MRR	7466.06 \pm 2271.20	6866.39 \pm 2958.28	7472.22 \pm 2267.48	6922.33 \pm 2905.46	7181.76
Qwen3 4B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	List-CoT	48.57 \pm 52.17	38.06 \pm 64.98	54.02 \pm 584.60	399.98 \pm 637.79	452.76
Qwen3 4B: RFT-QA	QA	QA-CoT	QA-Acc	List-CoT	564.63 \pm 489.19	470.46 \pm 698.12	618.16 \pm 522.63	440.12 \pm 440.77	523.34
Qwen3 4B: RFT-List (Acc)	QA	QA-CoT	QA-Acc	List-Acc	809.16 \pm 874.48	7734.09 \pm 1831.24	8046.00 \pm 1057.95	7841.95 \pm 1616.59	7928.64
Qwen3 4B: RFT-List (MRR)	QA	QA-CoT	QA-Acc	List-MRR	627.65 \pm 563.57	562.06 \pm 805.94	732.16 \pm 857.50	481.87 \pm 413.29	600.91
OpenThinker3 7B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	List-CoT	1106.96 \pm 1045.86	930.15 \pm 932.33	1181.00 \pm 1017.38	947.84 \pm 805.67	1041.49
OpenThinker3 7B: RFT-QA	QA	QA-CoT	QA-Acc	List-CoT	815.42 \pm 376.80	729.04 \pm 484.47	809.92 \pm 453.05	756.26 \pm 552.82	790.56
OpenThinker3 7B: RFT-List (Acc)	QA	QA-CoT	QA-Acc	List	67.37 \pm 12.09	62.09 \pm 12.99	69.00 \pm 12.09	60.00 \pm 10.19	68.00
OpenThinker3 7B: RFT-List (MRR)	QA	QA-CoT	QA-Acc	List-MRR	810.69 \pm 771.38	792.78 \pm 1247.39	810.57 \pm 767.94	8006.78 \pm 1145.71	8047.88
Qwen2.5 3B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	List-CoT	112.62 \pm 91.79	99.85 \pm 585.58	1250.13 \pm 876.76	988.88 \pm 849.23	1090.37
Qwen2.5 3B: RFT-QA	QA	QA-CoT	QA-Acc	List-CoT	835.24 \pm 1121.29	687.59 \pm 880.35	866.42 \pm 1052.75	603.96 \pm 688.68	748.30
Qwen2.5 3B: RFT-List (Acc)	QA	QA-CoT	QA-Acc	List-Acc	810.06 \pm 626.72	815.32 \pm 464.26	8173.18 \pm 300.29	8175.90 \pm 308.29	8157.87
Qwen2.5 3B: RFT-List (MRR)	QA	QA-CoT	QA-Acc	List-MRR	7638.32 \pm 1963.90	7596.11 \pm 2086.31	7819.54 \pm 1627.80	7685.50 \pm 1898.64	7678.12
AlphaMed 7B: RFT-MCQ	MCQ	MCQ-CoT	MCQ-Acc	List-CoT	332.31 \pm 778.99	215.31 \pm 459.06	319.19 \pm 766.59	252.24 \pm 110.00	304.21 \pm 155.02
AlphaMed 7B: RFT-QA	QA	QA-CoT	QA-Acc	List-CoT	267.25 \pm 408.98	170.93 \pm 251.13	220.24 \pm 101.00	183.21 \pm 155.02	219.13
AlphaMed 7B: RFT-List (Acc)	QA	QA-CoT	QA-Acc	List	269.51 \pm 335.98	46.00 \pm 100.01	254.44 \pm 358.09	304.21 \pm 305.24	318.16
AlphaMed 7B: RFT-List (MRR)	QA	QA-CoT	QA-Acc	List-MRR	204.34 \pm 106.99	157.38 \pm 74.03	261.13 \pm 101.89	188.60 \pm 94.62	217.86
RFT-List (List-CoT, MRR, LP=0)	QA	List-CoT	List-MRR-0	List-CoT	150.30 \pm 325.95	119.60 \pm 175.40	137.81 \pm 31.18	119.40 \pm 30.37	131.78
RFT-List (List-CoT, MRR, LP=0.1)	QA	List-CoT	List-MRR-0.1	List-CoT	131.40 \pm 185.45	114.02 \pm 203.35	142.15 \pm 4.40	120.06 \pm 12.06	124.06
RFT-List (List-CoT, MRR, LP=0.3)	QA	List-CoT	List-MRR-0.3	List-CoT	114.34 \pm 5.52	10.89 \pm 4.47	11.24 \pm 6.61	10.34 \pm 6.60	10.96
RFT-List (List-CoT, MRR, LP=0.5)	QA	List-CoT	List-MRR-0.5	List-CoT	10.27 \pm 2.49	9.74 \pm 2.25	10.20 \pm 2.81	9.58 \pm 2.38	9.95
RFT-List (List-CoT, MRR, LP=0.7)	QA	List-CoT	List-MRR-0.7	List-CoT	10.63 \pm 2.88	10.46 \pm 3.67	10.80 \pm 3.59	10.13 \pm 3.10	10.51
RFT-List (List-CoT, MRR, LP=0.9)	QA	List-CoT	List-MRR-0.9	List-CoT	10.67 \pm 2.82	10.27 \pm 3.07	10.65 \pm 3.29	10.13 \pm 3.11	10.43
AlphaMed 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	List-CoT	178.32 \pm 329.44	122.20 \pm 497.45	167.40 \pm 64.77	134.74 \pm 58.33	150.66
AlphaMed 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	List-CoT	772.42 \pm 442.14	649.14 \pm 457.05	882.74 \pm 621.16	656.73 \pm 356.51	741.51
OpenThinker3 7B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	List-CoT	752.86 \pm 521.78	660.12 \pm 418.33	785.64 \pm 467.11	681.26 \pm 508.39	719.97
Qwen2.5 3B: RFT-List (Acc, LP=0.3)	QA	List-CoT	List-Acc-LP-0.3	List-CoT	139.21 \pm 35.86	117.15 \pm 175.88	145.37 \pm 179.86	119.65 \pm 35.07	130.34

2889

Table 36: Metrics related to the ranked list answer format from the generated evaluation responses across benchmarks from Section 4.2.

2890

	Dataset	Prior Prompt	Reward Function	Prompting	MedQA	MedMCQA	MXOA	MLUP-H				
				LL	VLL	CP	LL	VLL	CP	LL	VLL	CP
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	List	501.39 \pm 45.49	658.82 \pm 567.17	656.77 \pm 67.14	643.31 \pm 643.31	5.15			
RFT-List (Baseline, List, Acc)	QA	List	List-Acc	List-CoT	243.62 \pm 204.28	28.36 \pm 36.64	232.58 \pm 232.92	4.6	263.06 \pm 263.06	2.47		
RFT-List (List-CoT, Acc)	QA	List	List-CoT	List	9.55 \pm 5.55	5.78 \pm 5.79	5.97 \pm 5.37	5.03 \pm 6.03	1.73			
RFT-List (List, MRR)	QA	List	List-MRR	List	9.55 \pm 2.36	9.60 \pm 2.75	9.39 \pm 9.39	3.08 \pm 9.49	9.49 \pm 2.17			
RFT-List (List, MRR)	QA	List	List-MRR	List-CoT	6.70 \pm 6.71	6.81 \pm 6.31	6.31 \pm 6.97	6.98 \pm 2.36	8.02 \pm 8.03	1.87		
RFT-List (List-CoT, MRR)	QA	List	List-CoT	List	21.90 \pm 1.89	15.13 \pm 2.31	17.93 \pm 17.94	2.41 \pm 12.90	12.90 \pm 12.90	1.82		
RFT-List (List-CoT, Judge-MRR)	QA	List	List-CoT	LLM-List-MRR	4.60 \pm 4.61	1.54 \pm 4.29	4.29 \pm 4.57	1.84 \pm 4.36	4.37 \pm 1.39			
RFT-MCQ (Baseline, No Prompt)	MCQ	No Prompt	MCQ-Acc	List-CoT	2.46 \pm 2.46	1.29 \pm 1.89	1.89 \pm 1.34	2.42 \pm 2.43	1.49 \pm 2.14	2.14 \pm 1.23		
RFT-MCQ (Baseline, MCQ)	MCQ	MCQ	MCQ-Acc	List-CoT	2.34 \pm 2.34	1.40 \pm 1.83	1.83 \pm 1.37	2.32 \pm 2.33	1.61 \pm 1.61	1.62 \pm 1.11		
RFT-MCQ (Baseline, MCQ)	MCQ	MCQ	MCQ-Acc	List-CoT	1.52 \pm 1.57	1.17 \pm 1.44	1.45 \pm 1.17	1.60 \pm 1.64	1.30 \pm 1.30	1.62 \pm 1.11		
RFT-MCQ (MCQ-CoT)	MCQ	MCQ	MCQ-Acc	List-CoT	2.46 \pm 2.46	1.43 \pm 1.40	2.07 \pm 2.07	2.44 \pm 2.44	1.59 \pm 1.59</			

2916 The attending tells the resident that the patient will do fine, and there is no need to report this
 2917 minor complication that will not harm the patient, as he does not want to make the patient worry
 2918 unnecessarily. He tells the resident to leave this complication out of the operative report. Which of
 2919 the following is the correct next action for the resident to take?" The available choices are omitted
 2920 for brevity. The ground-truth answer is "B) Tell the attending that he cannot fail to disclose this
 2921 mistake."

2922 Although not all responses for the MCQ task exhibit this behavior, it occurs frequently enough to
 2923 cause the model to miss points due to formatting errors. This illustrates how SFT with synthetic
 2924 reasoning traces can lead the model to overfit to the specific training format, reducing its ability to
 2925 follow other instructions. In contrast, the response generated by m1 for the same question, shown in
 2926 Figure 29, follows the formatting instruction correctly.

2927

2928

2929

2930 ## Thinking

2931

2932 Okay, so the resident accidentally cuts a flexor tendon during the carpal tunnel repair. That's
 2933 definitely a mistake, but it's not causing any harm to the patient. The attending says it's fine and tells
 2934 the resident not to mention it in the report. Hmm, that's a bit tricky. On one hand, the attending is
 2935 trying to protect the patient from unnecessary worry, which is kind of nice. But on the other hand, isn
 2936 't it important to be completely transparent with the patient about what happened during their surgery
 2937 ? I mean, the patient has a right to know everything that went on, even if it's a small complication.

2938

2939

2940

2941 Now, let's think about the ethical side of this. The resident is in a tough spot. The attending is
 2942 suggesting they leave it out of the report, but that feels like it could be misleading. The resident has to
 2943 weigh the importance of honesty with the patient against the attending's advice. It's a balancing act.

2944

2945

2946

2947

2948

2949

2950

2951

2952

2953

2954

2955

2956

2957

2958

2959

2960

2961

2962

2963

2964

2965 ## Final Response

2966

2967

2968

2969

2970 A) Disclose the error to the patient and put it in the operative report

2971 The resident should disclose the error to the patient and include it in the operative report. This action
 2972 ensures that the patient is fully informed about what happened during their surgery, respecting their
 2973 right to know all aspects of their medical care. Additionally, it maintains the integrity of the medical
 2974 records, which is crucial for both ethical and legal reasons.

Figure 28: Generated response from HuatuoGPT-o1 ignore the final answer instruction.

O.1 SFT MODELS

2966 Figures 30 to 32 show examples of generated responses to the question: 'What viral disease is the
 2967 only one to have been eradicated worldwide through vaccination efforts? The ground-truth answer
 2968 is "smallpox." The responses are produced by SFT-MCQ, SFT-QA, and SFT-List under MCQ, QA,
 2969 and List evaluations, respectively. Figure 33 shows a shortened response generated by SFT-MCQ
 during the List evaluation, containing repeated items within the generated list.

2970

2971

2972

2973

2974

2975

2976

2977

2978

2979

2980

2981

2982

2983

2984

2985

2986

2987

2988

2989

2990

2991

2992

2993

2994

2995

2996

A.

The correct next action for the resident is to **A) Disclose the error to the patient and put it in the operative report**.

Explanation:

Cutting a flexor tendon during carpal tunnel repair is a significant complication, even if it is repaired without further issues. Ethically and legally, the resident has a duty to transparently communicate this error to the patient. Withholding such information could breach trust and potentially lead to legal consequences. Additionally, the operative report is a critical document for patient care and legal records; omitting the complication would be inaccurate and could harm future care if the patient is unaware. The resident should report the complication honestly, ensuring transparency while providing necessary support to the patient.

Why Other Options Are Incorrect:

- **B)** Telling the attending they cannot omit the mistake is unprofessional and could escalate the situation without addressing the core issue.
- **C)** Reporting the physician to the ethics committee is premature and could damage the learning environment or lead to unnecessary conflict without verifying facts.
- **D)** Refusing to dictate the report is ineffective, as the resident is still responsible for accurate documentation.

\boxed{A}

Figure 29: Generated response from m1 that follows the formatting instruction.

O.2 RFT MODELS

Figures 34 to 38 show examples of generated responses from the RFT models for the same question: In the scenario where a surgical attending instructs a resident not to disclose an intraoperative complication to the patient or include it in the operative report, what is the most appropriate immediate action the resident should take?" The ground-truth answer is Tell the attending that he cannot fail to disclose this mistake." The figures correspond to responses from RFT-MCQ, RFT-QA, RFT-List-Acc, RFT-List-MRR, and RFT-List-Judge-MRR evaluated under their respective answer formats. Figure 39 shows an additional example from the RFT-QA model, which exhibits list-like behavior in its final answer. Figures 40 and 41 show examples of generated responses from RFT-List-Judge-MRR under MCQ and QA evaluations.

3005

3006

3007

3008

3009

3010

3011

3012

3013

3014

3015

3016

3017

3018

3019

3020

3021

3022

3023

3024
 3025
 3026
 3027
 3028 <think>
 Okay, let's tackle this question. The user is asking about what viruses have encouraged us to change our world, with options from A to H. I need to recall historical and current medical achievements related to viruses.
 3029
 3030 First, I remember that viruses have been a major cause of diseases, so efforts to combat them have led to significant advancements. Let me go through each option one by one.
 3031 Option A: Developed a universal vaccine for all types of flu. I think there's no universal flu vaccine yet. Flu vaccines are usually updated annually
 3032 because the virus changes, so a universal one isn't available. So A is probably not correct.
 3033 Option B: Eradicated polio. Polio was eradicated globally through vaccination efforts, right? The Global Polio Eradication Initiative led by the WHO
 3034 and others has been successful, and polio is now almost eradicated. So B might be a candidate.
 3035 Option C: Created a vaccine for the common cold. The common cold is caused by many different viruses, and there's no vaccine for it. There are some
 3036 experimental ones, but none are widely available or effective. So C is likely incorrect.
 3037 Option D: Found a cure for all types of cancer. Cancer is a complex disease, and while there are treatments, there's no universal cure. Vaccines exist for
 3038 some cancers (like HPV for cervical cancer), but not for all types. So D is probably not right.
 3039 Option E: Eradicated smallpox. Smallpox was eradicated in 1980, thanks to the smallpox vaccine. That's a huge achievement. So E could be correct.
 3040 Option F: Completely stopped the spread of all infectious diseases. That's impossible because new diseases emerge, and some are always spreading. So
 3041 F is definitely wrong.
 3042 Option G: Developed immunity to all types of viruses. Immunity is specific to each virus, and we can't develop immunity to all viruses. Vaccines target
 3043 specific pathogens, so G is incorrect.
 3044 Option H: Eradicated HIV/AIDS. HIV/AIDS is still a major global health issue, and while there's progress, it's not eradicated. There's no cure, and it's
 3045 not eradicated yet. So H is not correct.
 3046 Now, between B and E. Polio eradication is ongoing, but it's been declared eradicated in many regions. Smallpox was eradicated globally in 1980. The
 3047 question says "eradicated," which means completely eliminated from the planet. Smallpox is the only human disease eradicated by vaccination. Polio is
 3048 close but not fully eradicated yet. The WHO has declared it eradicated in the wild, but there are still cases in a few countries. However, smallpox was
 3049 eradicated in 1980, which is a significant milestone.
 3050 Wait, the question says "eradicated polio." If polio is still not fully eradicated, then B might not be correct. But the question might be considering the
 3051 fact that polio is on the verge of eradication. However, the standard answer for such questions usually points to smallpox as the one that was eradicated.
 3052 Let me double-check.
 3053 Smallpox was eradicated in 1980, and polio is still being eradicated. The Global Polio Eradication Initiative is working towards that goal, but it's not yet
 3054 complete. So the correct answer should be E, eradicated smallpox.
 3055 Wait, but the question says "eradicated polio" as option B. If polio isn't fully eradicated, then B is incorrect. But the question might be using "eradicated"
 3056 "loosely, but in medical terms, eradication means global elimination. Smallpox is the only one that's been eradicated. So E is correct.
 3057 Let me confirm: Smallpox was eradicated in 1980, polio is still not eradicated. So the answer should be E.
 3058 So the correct answer is E. Eradicated smallpox.
 3059 </think>
 3060 The question asks what viruses have encouraged us to change our world, with the options provided. Let's evaluate each choice based on historical and
 3061 current medical achievements:
 3062 - **Option A (Developed a universal vaccine for all types of flu)**: Incorrect. Influenza vaccines are updated annually due to viral mutations, and a
 3063 universal vaccine for all flu strains does not exist.
 3064 - **Option B (Eradicated polio)**: Incorrect. While polio eradication efforts have been successful in many regions, polio is **not yet fully eradicated
 3065 ** globally. The Global Polio Eradication Initiative continues to work toward this goal, but it is not yet complete.
 3066 - **Option C (Created a vaccine for the common cold)**: Incorrect. The common cold is caused by multiple viruses (e.g., rhinoviruses), and no
 3067 effective vaccine exists for it.
 3068 - **Option D (Found a cure for all types of cancer)**: Incorrect. Cancer is a complex disease with no universal cure. While targeted treatments exist for
 3069 specific cancers, a cure for all types does not exist.
 3070 - **Option E (Eradicated smallpox)**: **Correct**. Smallpox was the **only human disease eradicated by vaccination**. The World Health
 3071 Organization (WHO) declared smallpox eradicated in 1980, following a global vaccination campaign. This is a landmark achievement in public health.
 3072 - **Option F (Completely stopped the spread of all infectious diseases)**: Incorrect. New infectious diseases emerge (e.g., SARS-CoV-2), and
 3073 diseases like HIV/AIDS persist, so this is impossible.
 3074 - **Option G (Developed immunity to all types of viruses)**: Incorrect. Immunity is specific to each virus; vaccines target particular pathogens, not all
 3075 viruses.
 3076 - **Option H (Eradicated HIV/AIDS)**: Incorrect. HIV/AIDS remains a major global health challenge with no cure, though antiretroviral therapy has
 3077 improved outcomes.
 3078 The key historical milestone is the eradication of smallpox, which is a direct result of viral disease control efforts. While polio eradication is ongoing,
 3079 smallpox is the only disease eradicated by vaccination.
 3080 \boxed{E}

Figure 30: Generated response from SFT-MCQ on MCQ evaluation.

3078
3079
3080
3081
3082
3083

3084
3085

<think>

3086 Okay, the user is asking about a viral disease that's been eradicated worldwide through vaccination
3087 efforts. Let me recall what I know about global health achievements.

3088

3089 Hmm, I remember that smallpox was the only disease eradicated through vaccination. The smallpox
3090 vaccine was developed by Edward Jenner in the 18th century, and after decades of global vaccination
3091 campaigns, the World Health Organization declared smallpox eradicated in 1980.

3092

3093 Wait, let me double--check to make sure I'm not confusing it with other diseases. Polio is close--
3094 there's been significant progress with the vaccine, but it's not fully eradicated yet. Measles and
3095 rubella are controlled but not eradicated. Rabies can be prevented with vaccines, but it's not
3096 eradicated globally.

3097

3098 The key point here is "eradicated worldwide," meaning no cases anywhere on Earth. Smallpox fits
3099 that description perfectly. The WHO certified it as eradicated in 1980, and since then, no natural
3100 cases have been reported.

3101

3102 I should confirm if there are any other candidates. Bacterial diseases like cholera or tuberculosis have
3103 been controlled with vaccines, but they're not eradicated. So smallpox is definitely the answer.

3104

3105 The user seems to want a straightforward, factual answer--probably for educational purposes. They
3106 specified the reasoning process in tags, so I'll make sure to follow that format exactly. No need for
3107 extra fluff; just the facts.

3108

3109 Final check: Yes, smallpox is the only viral disease eradicated through vaccination. The answer is
3110 clear-cut.
</think>

3111

3112 The question asks for the viral disease eradicated worldwide through vaccination efforts. To
3113 determine this, I recall historical and epidemiological facts about global disease eradication.

3114

3115 – Smallpox (caused by **Variola major** virus) was the only viral disease eradicated globally via
3116 vaccination. The smallpox vaccine, developed by Edward Jenner in 1796, was pivotal in the
3117 eradication campaign.

3118

3119 – The World Health Organization (WHO) launched a global eradication initiative in 1967, and after
3120 extensive vaccination efforts, the last naturally occurring case was reported in Somalia in 1977. The
3121 WHO declared smallpox eradicated in 1980.

3122

3123 – Other viral diseases like polio, measles, and rubella are controlled but not yet eradicated (polio
3124 remains endemic in a few countries, and measles has seen recent outbreaks despite vaccination).
3125 Bacterial diseases (e.g., cholera, tuberculosis) are not viral and thus irrelevant here.

3126

3127 No other viral disease has achieved complete global eradication through vaccination. Therefore,
3128 smallpox is the correct answer.

3129

3130 \boxed{\text{smallpox}}

3131

3125 Figure 31: Generated response from SFT-QA on QA evaluation.
3126
3127
3128
3129
3130
3131

```

3132
3133
3134 <think>
3135 Okay, the user is asking about a viral disease that has been eradicated worldwide through vaccination
3136 efforts. Let me recall what I know about disease eradication.
3137
3138 First, I remember that smallpox was the first and only disease eradicated globally. The World Health
3139 Organization (WHO) declared smallpox eradicated in 1980 after a massive vaccination campaign.
3140
3141 I should double-check if there are any other diseases that might have been eradicated. Polio is close
3142 to eradication but hasn't been officially declared eradicated yet. Measles and rubella are controlled
3143 but not eradicated. Rabies can be controlled with vaccination but isn't eradicated globally.
3144
3145 So, smallpox is the correct answer. The user might be a student or someone interested in public
3146 health, so accuracy is important. They probably need this for a test or to understand historical
3147 achievements in medicine.
3148
3149 I should confirm that smallpox is indeed the only one. Yes, WHO states that smallpox is the only
3150 human disease eradicated through vaccination. The answer should be straightforward.
3151 </think>
3152
3153 # Final Answer
3154 1. Smallpox
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185

```

Figure 32: Generated response from SFT-List on List evaluation.

```

# Final Answer
1. Tryptophan
2. Histidine
3. Methionine
4. Lysine
5. Leucine
6. Isoleucine
7. Valine
8. Threonine
9. Phenylalanine
10. Arginine
11. Alanine
12. Glycine
13. Proline
14. Serine
15. Cysteine
16. Tyrosine
17. Aspartic acid
18. Glutamic acid
19. Asparagine
20. Glutamine
21. Methionine
22. Cysteine
23. Tyrosine
// ...
// Omitted for brevity
// ...
1045. Cysteine
1046. Tyrosine
1047. Phenylalanine

```

Figure 33: Generated response from SFT-MCQ on List evaluation with repeating content.

3186

3187

3188

3189

<think>

The resident is in a situation where there is a minor complication that occurred during the surgery, specifically the inadvertent cut to a flexor tendon. The attending physician is suggesting that this complication be omitted from the operative report and not disclosed to the patient. However, it is important to adhere to ethical standards and patient safety protocols. Minor complications should still be documented and disclosed to ensure transparency and proper patient care.

3194

3195

3196

The resident has a duty to report the complication to the patient and include it in the operative report, even if the attending physician advises against it. This ensures that the patient is informed and that the medical record is accurate.

3197

3198

Options:

- A) Disclose the error to the patient and put it in the operative report – This is the correct action as it adheres to ethical standards and patient rights.
- B) Tell the attending that he cannot fail to disclose this mistake – This is also correct but does not address the reporting to the patient.
- C) Report the physician to the ethics committee – This is an extreme measure and not necessary in this situation.
- D) Refuse to dictate the operative report – This is not the best approach as it does not address the patient's right to know and the need for accurate documentation.

The correct action is to disclose the error to the patient and include it in the operative report.

</think>

\boxed{A}

3209

3210

Figure 34: Generated response from RFT-MCQ on MCQ evaluation.

<think>

The scenario presents an ethical dilemma involving patient safety, transparency, and legal compliance. The resident is instructed not to disclose an intraoperative complication to the patient or include it in the operative report. This instruction is unethical and potentially illegal, as patients have a right to know about complications that occur during surgery, and failure to report complications in the operative record can lead to malpractice lawsuits and harm to patient care in the future.

The resident should not follow this instruction and should take immediate action to ensure patient safety and compliance with medical ethics and legal standards. The resident should:

1. Document the complication in the patient's medical record, ensuring it is included in the operative report.
2. Inform the patient about the complication as soon as it is safe to do so, respecting the patient's right to know and the principle of informed consent.
3. Report the incident to a supervisor or the hospital's risk management department to address the ethical and legal concerns.
4. Consider seeking advice from a legal or ethical consultant within the hospital to ensure compliance with all relevant standards and laws.

</think>

The most appropriate immediate action the resident should take is to \boxed{document the complication in the operative report and patient's medical record, inform the patient about the complication as soon as it is safe to do so, report the incident to a supervisor or the hospital's risk management department, and seek advice from a legal or ethical consultant if necessary}.

3236

3237

3238

3239

Figure 35: Generated response from RFT-QA on QA evaluation.

3240
3241
3242
3243
3244
3245
3246
3247

<think>In this scenario, the resident is faced with a directive that conflicts with ethical standards and legal requirements for patient safety and transparency. The resident should not follow the attending's instruction to withhold information from the patient or omit it from the operative report. The appropriate actions include reporting the situation to a supervisor, the department head, or the institution's ethics committee. The resident should also document the incident in their own notes and consider seeking advice from a more senior colleague or legal counsel.</think>

3248
3249
3250
3251
3252
3253
3254
3255

Final Answer

1. Report the incident to a senior surgeon
2. Document the incident in the resident's notes
3. Seek advice from a more senior colleague
4. Report the incident to the department head
5. Report the incident to the institution's ethics committee
6. Follow the attending's instruction

3256
3257
3258
3259
3260

<think>The resident is faced with a ethical dilemma involving patient safety and transparency. The most appropriate action would be to report the complication to a supervisor or the medical ethics committee to ensure patient safety and maintain proper documentation. The resident should not follow the attending's instruction to withhold information from the patient or omit it from the report .</think>

3266
3267
3268
3269
3270
3271
3272

Final Answer

1. Report the complication to a senior surgeon
2. Report the complication to the medical ethics committee
3. Follow the attending's instructions
4. Disclose the complication to the patient directly
5. Document the complication in the patient's chart without reporting it

3273
3274
3275
3276
3277

Figure 36: Generated response from RFT-List-Acc on List evaluation.

3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291

<think>Considering the ethical and legal responsibilities of a medical professional, the resident should prioritize patient safety and transparency. The options should include reporting the issue to a supervisor, the medical director, or a hospital ethics committee if the attending's instructions are to withhold information from the patient and not document it. The most appropriate action is to report the attending's instructions to a higher authority or ethics committee to ensure patient safety and compliance with medical ethics.</think>

Final Answer

1. Report the attending's instructions to a hospital ethics committee
2. Report the attending's instructions to the medical director
3. Follow the attending's instructions without reporting them
4. Discuss the situation with a senior colleague or another attending physician
5. Document the complication in the operative report and inform the patient without informing the attending

3292
3293

Figure 37: Generated response from RFT-List-MRR on List evaluation.

3294

3295

3296

3297

3298

<think>

Copper IUDs are generally considered safe and effective for most women. However, there are certain medical conditions that make their use contraindicated. These include:

1. Pelvic inflammatory disease (PID) or a history of PID, as the IUD can increase the risk of infection

.

2. Active genital or pelvic infection at the time of insertion.

3. Uterine or cervical cancer.

4. Known or suspected pregnancy.

5. Severe liver disease.

6. Allergy to copper or other IUD components.

7. Certain sexually transmitted infections (STIs) or risk factors for STIs.

8. Previous uterine surgery that has altered the uterine cavity.

9. Adhesions from previous pelvic surgery.

10. Severe endometriosis or adenomyosis, as the IUD can exacerbate symptoms.

11. Copper IUD is contraindicated in women with Wilson's disease, as copper toxicity can be a concern.

12. Thrombophlebitis or thrombosis, as the IUD can increase the risk of thromboembolic events.

13. Active genital herpes or other sexually transmitted infections that could be exacerbated by the IUD.

Given the question, we are specifically looking for past medical history that would make copper IUD placement contraindicated. This would include a history of pelvic inflammatory disease, previous pelvic or uterine surgery, endometriosis, adenomyosis, or Wilson's disease.

</think>

The past medical history that would make copper IUD placement contraindicated in a woman requesting this form of contraception includes \boxed{pelvic inflammatory disease, previous pelvic or uterine surgery, endometriosis, adenomyosis, Wilson's disease, or a history of severe liver disease, active genital or pelvic infection, or known or suspected pregnancy at the time of consultation}.

3317

3318

3319

3320

3321

3322

3323

Figure 39: Generated response from RFT-QA for the question: “What past medical history would make copper intrauterine device (IUD) placement contraindicated in a woman requesting this form of contraception?” The ground-truth answer is “Active or recurrent pelvic inflammatory disease (PID).”

3324

3325

3326

3327

3328

3329

3330

3331

3332

3333

<think> The resident has a duty to ensure that the patient is fully informed about the procedures and any complications that occur during surgery. Additionally, accurate and complete documentation is crucial for medical records and patient care. The attending's suggestion to omit the complication from the operative report and not inform the patient is unethical and could lead to potential harm if the patient is unaware of the issue. The resident has a professional and ethical obligation to report the complication and include it in the operative report.</think>

The correct next action for the resident to take is to disclose the error to the patient and put it in the operative report.

\boxed{A}

3343

3344

3345

3346

3347

Figure 40: Generated response from RFT-List-Judge-MRR on MCQ evaluation.

3348

3349

3350

3351

3352

3353

3354

3355

3356

3357

3358

3359

3360

3361

3362

3363

3364

3365

3366

3367

3368

3369

<think>Considering the ethical and legal implications, the resident should first question the attending to understand the reasoning behind the instruction. However, if the attending insists on not disclosing the complication to the patient and not including it in the operative report, the resident should report this to a more senior physician or the hospital's ethics committee to ensure patient safety and compliance with medical standards. Directly following the attending's instructions without question could potentially compromise patient care and legal standards.</think>

The most appropriate immediate action the resident should take is to report the situation to a more senior physician or the hospital's ethics committee.

\boxed{Report the situation to a more senior physician or the hospital's ethics committee}

3380

3381

Figure 41: Generated response from RFT-List-Judge-MRR on QA evaluation.

3382

3383

3384

3385

3386

3387

3388

3389

3390

3391

3392

3393

3394

3395

3396

3397

3398

3399

3400

3401