

Meerkat2: Bridging Global Reasoning and Local Knowledge for Efficient Medical AI

Jungwoo Park^{1,2}, Hyeongsoon Hwang^{1,2}, Sanghoon Lee^{1,2*}, Jaewoo Kang^{1,2*}

¹Korea University, ²AIGEN Sciences

{jungwoo-park, hhs8746}@korea.ac.kr

{sanghoon.lee, jaewoo.kang}@aigensciences.com

Abstract

Recent large language models are increasingly being adapted for medical reasoning, yet most open medical LLMs remain optimized for English-centric clinical settings and fail to capture local legal and institutional knowledge. We present **Meerkat2**, a family of open medical reasoning models at 14B, 32B, and 30B-A3B scales for Korean and English clinical contexts. Meerkat2 is trained on an 837K-example mixture that separates reasoning transfer from high-fidelity local knowledge: clinical reasoning is distilled from strong teachers, while Korean medical statutes are converted into source-grounded question-answer pairs. We also introduce **Verbosity-Decoupled Distillation (VDD)**, which preserves the benefits of long teacher rationales without making verbose generation to the student’s default behavior. Across MedQA and KorMedMCQA, Meerkat2 substantially improves over comparable base models, with the 30B-A3B variant reaching Korean licensing accuracy that matches Qwen3-235B-A22B while using only 3B active parameters. Ablations show that statute-grounded synthesis outperforms standard distillation, and mechanistic analysis reveals that VDD cancels gradients toward verbose generation. These results provide a practical recipe for localized and efficient medical LLMs.¹

1 Introduction

Large language models (LLMs) are increasingly proposed as biomedical assistants, supporting clinical decision-making, patient communication, and access to medical information (Thirunavukarasu et al., 2023; Tian et al., 2024). The sensitivity of patient data and concerns about medical privacy have further driven demand for open models that can be deployed on hospital premises (Wiest et al.,

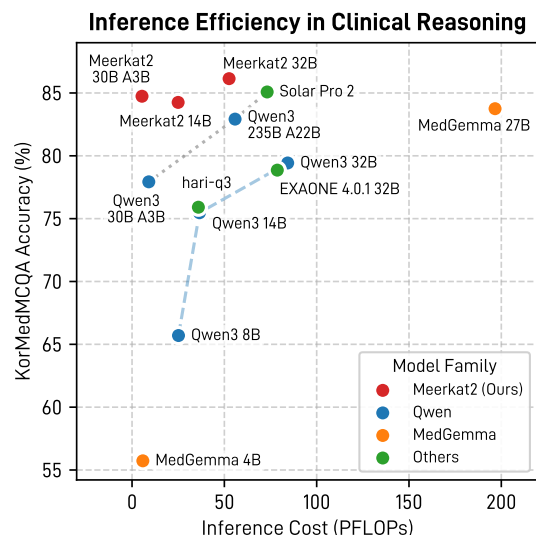


Figure 1: Trade-off between inference cost and KorMedMCQA accuracy. The x-axis denotes the total inference cost used for evaluation. Meerkat2 shows a more favorable trade-off across other models.

2024; Zhang et al., 2023; Kim et al., 2025). Several practical obstacles, however, limit how far medical LLMs can be pushed under real-world constraints.

Clinical practice, pharmacology, and regulation closely depend on the country in which they are applied, and English-centric medical LLMs do not transfer well to other healthcare environments (Zhang et al., 2023; Kweon et al., 2024; Qiu et al., 2024). Adapting a model to a local environment is not a matter of translation alone. Each country has its own clinical guidelines and medical laws, and these must be reflected in the training data rather than approximated through a teacher. They depend on accurate memorization and require high fidelity, which teacher distillation does not guarantee. A separate concern arises in the distillation of reasoning itself. Long chains of thought (Wei et al., 2022; Chen et al., 2025a; Yun et al., 2025) are effective for transferring reasoning ability (Hugging Face, 2025a; Guha et al., 2025), but the student tends to imitate the teacher’s verbosity, generating long outputs even when a short answer is enough. Even

* Corresponding authors

¹Models are available at <https://huggingface.co/collections/aigensciences/meerkat2>

when the per-query delay is modest, unnecessary verbosity imposes a recurring interaction cost in clinical decision-support workflows, where effective systems must fit into clinician workflow and provide support at the time and location of decision making (Kawamoto et al., 2005).

We present **Meerkat2**, a family of open medical reasoning models at 14B, 32B, and 30B-A3B (MoE) scales, trained on a 837K-example mixture covering clinical reasoning, Korean medical statutes, biomedical question answering, and general-capability replay. Clinical reasoning is distilled from strong teachers. Statutory text requires verbatim fidelity, and question-answer pairs are synthesized directly from the source text. To handle the verbose response style that the student inherits during reasoning distillation, we introduce **Verbosity-Decoupled Distillation (VDD)**, which pairs long teacher rationales with an explicit verbosity instruction in the system prompt. We further analyze this effect through mechanistic interpretability, finding that the verbosity instruction cancels the gradient that would otherwise push the model toward verbose generation.

Across MedQA and KorMedMCQA, Meerkat2 substantially improves over comparable base models, and on the KorMedMCQA Doctor subset 30B-A3B scale reaches the competitive accuracy with only 3B active parameters. The same pipeline transfers to other backbones, with our Gemma3-4B (Gemma Team, 2025) model outperforming MedGemma-4B (Sellergren et al., 2025), a medically specialized model fine-tuned from the same base. Gains also generalize to the 89th Korean Medical Licensing Examination, which postdates our training cutoff. They further extend beyond examinations to clinical case diagnosis, information seeking, and open-ended clinical conversation.

Our contributions are as follows:

- We release Meerkat2, an open medical reasoning model family at 14B, 32B, and 30B-A3B scales, together with a 837K training mixture spanning clinical reasoning, Korean medical statutes, biomedical question answering, and general capabilities.
- We introduce Verbosity-Decoupled Distillation, which preserves the reasoning gains of long teacher rationales without internalizing verbose style, and explain its mechanism by showing that the verbosity instruction cancels the gradient toward verbose generation.

- We show that direct synthesis of question-answer pairs from Korean medical statutes substantially outperforms teacher distillation on the legal subset of KorMedMCQA, confirming the necessity of source-grounded supervision for knowledge requiring high fidelity.

2 Related Work

Reasoning in LLMs The reasoning capability of LLMs was initially explored through Chain-of-Thought (CoT) prompting (Wei et al., 2022). The introduction of OpenAI’s o1 (OpenAI, 2024) triggered the consideration of advanced reasoning, leading to the development of reasoning specialized models such as QwQ (Qwen Team, 2025b) or DeepSeek-R1 (Guo et al., 2025).

These models are typically trained with reinforcement learning (Shao et al., 2024) or distilled from teacher models (Kim and Rush, 2016). Several data collections such as OpenR1 (Hugging Face, 2025a), OpenMathReasoning (Moshkov et al., 2025), and OpenThoughts3 (Guha et al., 2025) have shown that curating high-quality reasoning trajectories is necessary to improve the reasoning capabilities of small language models. Yue et al. (2025) also observed that distillation is more effective at injecting new knowledge-intensive reasoning patterns than reinforcement learning.

Medical Reasoning While most reasoning abilities of LLMs are aimed at math and coding domains, there is growing evidence that explicit reasoning is also beneficial in biomedical and clinical domains. Meerkat (Kim et al., 2025) achieved significant performance on USMLE with high-quality CoTs and medical textbook contents. HuatuoGPT-o1 (Chen et al., 2025a) and Med-RLVR (Zhang et al., 2025) confirmed that advanced reasoning is also effective in the biomedical domain. Med-PRM (Yun et al., 2025) employed a Process-Reward Model (PRM) to evaluate each step of clinical reasoning and achieved performance gains, demonstrating the importance of reasoning in biomedical domain. Meanwhile, Chen et al. (2023) and Qiu et al. (2025) remarked that domain specialization can harm robustness and lead to catastrophic forgetting.

3 Data Curation

We organize the training data into 4 categories and curate each category carefully. Our objective is to develop a chatbot that performs effectively in both

Case Study: Clinical Reasoning Quality across CoT Variants (KorMedMCQA Doctor Subset)		
Question A 48-year-old man presents with 3 hours of epigastric pain. Pain started 6 hours after eating raw saltwater fish (sashimi). Intermittent squeezing/cramping epigastric pain with nausea and vomiting. BP 135/80, HR 95, RR 18, T 36.5°C. Acutely ill-appearing; epigastric tenderness, no rebound. Blood: Hb 13.1, WBC 8,300, Plt 290k, T.bili 1.2, AST 39, ALT 34, ALP 130, lipase 21. Stool: no WBCs, no occult blood. Choices: A) Upper endoscopy B) Capsule endoscopy C) Colonoscopy D) Stool culture E) Clostridium toxin assay Ground truth: A		
Distilled from Concise CoTs (Frequent clinical-reasoning errors; 690 Tokens) Epigastric pain 6 h after raw sashimi, N/V, no fever, labs/stool normal — so acute gastroenteritis (esp. toxin-mediated food poisoning) should be suspected. • A) Upper endoscopy — for upper-GI bleed, ulcer, tumor. Acute food poisoning — not indicated • B) Capsule endoscopy — for small-bowel disease. Does not fit acute food poisoning. • C) Colonoscopy — for lower-GI disease. Epigastric symptoms, stool clean — not indicated • D) Stool culture — for bacterial enteritis. But here, V/pain within hours of sashimi, no fever / no bloody stool / no fecal WBCs → classic toxin-mediated food poisoning (e.g. <i>Vibrio</i>) — not first test. • E) Clostridium toxin assay — diagnoses <i>C. perfringens</i> food poisoning (meat, 8–16 h, diarrhea-predominant). This case (raw fish, 6 h, V/pain) fits <i>Vibrio</i>-type toxin poisoning better — among the choices, only E is tied to toxin-mediated food poisoning. ✗ ANSWER: E	Distilled from Verbose CoTs (Correct, but with redundant filler; 1,285 Tokens) • 48 y/o; onset 6 h after raw sashimi; squeezing epigastric pain with N/V; vitals stable; epigastric tenderness without rebound (no peritonitis); WBC/LFTs/lipase normal; stool no WBCs/occult blood. - The first thing to consider is acute gastritis (food-induced). Epigastric pain, N/V within hours of food intake fit a food-poisoning / gastritis pattern well. A) Upper endoscopy • Considered first-line when food-induced acute gastritis is suspected — rule out bleeding, ulcer, malignancy. • Vitals stable, no peritoneal signs, bloods/stool bland — directly inspecting the gastric mucosa is reasonable ... upper endoscopy is the most appropriate test. B) Capsule endoscopy • Evaluates small-bowel lesions (IDA, SB bleed, Crohn's ...). • This patient has [...] no SB findings — not appropriate. C) Colonoscopy • For colon / lower-GI disease (colon cancer, polyps, IBD, hematochezia ...). • This patient has [...] no lower-GI findings — not appropriate. D) Stool culture • Performed when bacterial enteritis is suspected (diarrhea, fever, abd. pain ...). • This patient has [...] no diarrhea / fever / fecal WBCs — not appropriate. E) Clostridium toxin assay • <i>C. difficile</i> toxin assay — for antibiotic-associated diarrhea / pseudomembranous colitis. • This patient has ... no antibiotic exposure / diarrhea / fever — not appropriate. Summary — acute epigastric pain after food, no peritoneal signs ... → suspect acute gastritis ... most appropriate test is upper endoscopy. ✓ ANSWER: A	
Verbosity-Decoupled Distillation (Short, but clinically sound reasoning; 548 Tokens) Onset 6 h after raw sashimi, N/V, no fever, no diarrhea/bleeding, stool clean, labs normal — findings do not support bacterial enteritis or clostridial toxin food poisoning. • Bacterial enteritis: usually diarrhea (bloody/mucoid), fever, fecal WBCs → does not fit. • Clostridial toxin food poisoning: neurologic findings → absent. • Viral gastroenteritis: diarrhea predominates, but here pain is localized epigastric. • Acute pancreatitis: severe pain + elevated lipase — but lipase is normal. Most likely diagnosis: acute gastritis (food-induced). Test (with possible concurrent therapy): upper endoscopy. ✓ ANSWER: A		
■ Good reasoning ■ Wrong reasoning ■ Unnecessary verbosity		

Figure 2: Qualitative comparison on a KorMedMCQA Doctor question across three distillation variants, with reasoning steps evaluated and highlighted by a licensed physician. The Concise student reaches an incorrect answer, the Verbose student answers correctly with redundant restatement of findings, and the VDD student reaches the correct answer through compact reasoning, preserving clinical quality without inheriting verbose style.

U.S. and Korean clinical contexts while preserving the general-purpose capabilities.

3.1 Clinical Reasoning

First, we aim to improve clinical reasoning by distilling knowledge from powerful teacher models. Specifically, we use questions from MedQA (Jin et al., 2021) train set, KorMedMCQA (Kweon et al., 2024) train set, and MedBooks-18 (Kim et al., 2025) to synthesize clinical reasoning trajectories for problem solving. For the teachers, DeepSeek-V4-Pro (DeepSeek-AI, 2026) and Qwen3-235B-A22B-Thinking (Qwen Team, 2025a) are employed to generate CoTs and internal reasoning content. Each model attempts every question multiple times, and only correct responses are retained. For each instance, we randomly sample 6 different prompt templates (see Figure 4) to reduce dependence on the specific format presented in the training data.

Verbosity-Decoupled Distillation In our preliminary experiments, we observed that long and detailed CoTs are more effective for distillation than concise ones, leading to higher benchmark scores, which is consistent with Guha et al. (2025). However, we also found that student models tend to imitate the teacher’s verbosity.

More specifically, we controlled the verbosity

level with the system prompts and observed comparable performance across both concise and verbose responses. Nevertheless, student models distilled from concise outputs consistently achieved inferior performance. This indicates that while verbosity is not a prerequisite for correct reasoning at inference time, it is instrumental for achieving high-fidelity knowledge transfer during the distillation process.

Although verbose CoTs are beneficial for distillation, their response style can also be transferred to the student. Inspired by Kang et al. (2025), we include both concise and verbose CoTs in the training dataset, but attach an explicit verbosity instruction to the verbose examples through the system prompt, e.g., “Give me a very detailed, long-form answer.” Without this prompt, the student may interpret verbosity as the desired default response style and imitate it unconditionally. By explicitly conditioning verbose CoTs on a verbosity instruction, their length is attributed to the prompt constraint rather than to the reasoning trajectory itself. This allows us to retain the distillation benefits of detailed CoTs while preventing verbosity from being learned as the default response style.

3.2 Korean Medical Laws

Meanwhile, healthcare systems and clinical practice are closely related to legal and regulatory frameworks. In particular, legal knowledge largely

depends on accurate memorization and therefore requires high fidelity (Lee et al., 2025). This calls for data that explicitly preserves legal facts and their supporting rationale, rather than imitating teachers.

Unlike clinical reasoning, we found that distillation from teacher models is ineffective for medical legal knowledge, even though the teachers achieve high scores on the legal problems. We therefore synthesize explicit questions and solutions accompanied by detailed explanatory notes to facilitate memorization. These data are derived from 12 Korean statutes related to clinical practice (Table 8).

3.3 Biomedical Question Answering

Beyond specialized clinical reasoning, robust medical assistants should also support broad biomedical knowledge and open-ended medical conversation. Existing datasets such as MedQA and KorMedMCQA target medical licensing examinations and cover only a limited range of biomedical question types. Recent web-curated instruction datasets such as OpenThoughts (Guha et al., 2025), WebInstruct (Yue et al., 2024), and NaturalReasoning (Yuan et al., 2025) provide broader and more diverse supervision. We collect biomedical questions from these sources to improve coverage beyond exam-style reasoning. We first filter the questions for biomedical relevance using Qwen3-4B and then regenerate the answers with DeepSeek-V4-Pro, resulting in 57K conversations.

3.4 General LLM Capabilities

Finally, we incorporate general instruction-tuning data into the training mixture to mitigate catastrophic forgetting on general tasks. As noted in previous work (Rolnick et al., 2019; Scialom et al., 2022), continual fine-tuning on a specific domain can degrade the original abilities of an LLM. For example, since the three categories described above lack multi-turn conversations and are limited to English and Korean, the model risks losing its versatility. To retain these capabilities, we include a subset of general instruction data as replay examples.

Specifically, we curate subsets from SmolTalk2 (Hugging Face, 2025b) and Tulu-3 (Lambert et al., 2025), selecting five primary categories: (1) **System Prompt**, to ensure adherence to role-playing or persona constraints; (2) **Multilingual**, to maintain knowledge of languages present in the base model; (3) **Multi-Turn Conversation**, to preserve dialogue context handling; (4) **Instruction Following**, to execute general tasks such as rewriting and

summarizing; and (5) **Safety**, to prevent jailbreaking and enforce refusal of harmful requests. The detailed mixture is provided in Appendix A.1.

4 Experiments

4.1 Experimental Setup

To develop our medical reasoning models, we fine-tune three Qwen3 variants (14B, 32B, and 30B-A3B MoE) on the curated 837K dataset. Since Qwen3 supports hybrid thinking mode, we apply explicit token masking to the user prompts and empty reasoning blocks to discourage the model from unintentionally skipping the reasoning mode.

For our main evaluation of medical performance, we use the test sets of MedQA and KorMedMCQA, which assess clinical knowledge in U.S. and Korean contexts, respectively. We report performance in two modes: **Thinking**, where the model generates reasoning content within <think> tags, and **Instant**, where it answers directly. We also report MMLU (Hendrycks et al., 2021) and KMMLU (Son et al., 2025) as a preliminary assessment of general knowledge. We revisit advanced medical performance with a broader set of benchmarks in Section 4.3. All evaluations are conducted using EvalScope (ModelScope Team, 2024).

4.2 Main Results

Performance Gains Table 1 reports overall evaluation performance across the compared models. In **Instant** mode, Meerkat2 14B improves over Qwen3-14B by 9.2%p on MedQA and 7%p on KorMedMCQA. The larger variants show comparable gains: the 30B-A3B MoE model improves by 6.1%p and 5.9%p, and the 32B model by 8.3%p and 6.3%p, respectively. The advantage persists in **Thinking** mode, with the 14B model gaining 6%p on MedQA and 5.4%p on KorMedMCQA, and the 30B-A3B and 32B models showing improvements of 2.9%p to 5.2%p respectively.

KorMedMCQA Doctor Subset The Doctor subset, which is drawn from the Korean Medical Licensing Examination (KMLE), is a particularly demanding test of clinical competence in Korea. In **Instant** mode, Meerkat2-14B, 30B A3B, and 32B reach 86.4%, 88.5%, and 87.6%, respectively, which are competitive to Qwen3-235B-A22B (88.7%) despite using far fewer parameters. The 30B-A3B result is especially notable, as it matches frontier-scale accuracy with only 3B active parameters.

Model	MMLU	KMMLU	MedQA		KorMedMCQA		
			Instant	Thinking	Instant (English)	Instant (Korean)	Thinking
Small-Sized Models ($\leq 14B$)							
Qwen3-4B	76.9	58.3	69.0	78.1	67.5 (61.6)	64.7 (61.2)	74.1 (71.3)
MedGemma-4B	63.3	40.8	64.5	—	55.7 (56.3)	52.2 (45.5)	—
Qwen3-8B	77.3	58.6	72.7	80.8	65.7 (58.2)	65.2 (56.6)	76.3 (75.6)
Qwen3-14B	82.0	63.8	76.0	83.7	75.5 (75.6)	72.4 (67.1)	80.8 (81.6)
hari-q3-14B	80.8	61.2	75.3	85.8	75.9 (76.6)	73.2 (69.9)	82.6 (84.6)
Meerkat2-14B (Ours)	81.4	62.6	85.2	89.7	82.5 (86.4)	82.1 (88.5)	86.2 (90.6)
Base-Sized Models ($\leq 30B$)							
MedGemma-27B [†]	83.2	63.8	84.6	—	83.8 (86.2)	80.1 (81.8)	—
Qwen3-30B-A3B	83.1	65.2	81.2	85.5	77.9 (77.9)	74.7 (72.4)	81.4 (84.8)
Meerkat2-30B-A3B (Ours)	81.9	63.6	87.3	88.9	83.8 (88.5)	82.8 (87.6)	86.6 (92.6)
Medium-Sized Models ($\leq 32B$)							
Qwen3-32B	83.4	69.5	79.2	87.4	79.4 (82.1)	79.8 (79.3)	83.6 (86.4)
EXAONE 4.0.1 32B	85.9	69.9	81.5	88.1	78.9 (80.5)	76.9 (79.1)	85.1 (87.6)
Solar Pro 2	86.2	73.7	84.4	89.2	85.1 (87.8)	83.6 (87.8)	88.7 (91.7)
Meerkat2-32B (Ours)	83.9	67.1	87.5	90.3	85.7 (87.6)	85.0 (89.4)	88.2 (92.2)
Large-Sized Models ($> 32B$)							
Meerkat-70B	80.1	49.5	81.5	—	70.0 (68.3)	70.0 (69.0)	—
Qwen3-235B-A22B	88.6	75.9	89.8	92.9	86.7 (88.7)	88.0 (91.0)	90.8 (93.1)
GPT-5.1	90.9	77.5	93.2	96.4	91.0 (92.2)	90.8 (94.0)	95.1 (96.8)

Table 1: Performance of different models on various benchmarks. Values in parentheses indicate performance on **Doctor** subset in KorMedMCQA, which corresponds to Korean Medical Licensing Examination (KMLE). [†]The original paper reported performance of MedGemma-27B with test-time scaling and majority voting, while we report a vanilla run to match the evaluation setting used for other models.

Model	Diagnosis Arena	MediQ	HealthBench Hard	Avg.
Qwen3-14B	8.9	51.0	8.9	22.9
Meerkat2-14B	12.8	60.3	9.1	27.4
Qwen3-30B-A3B	12.0	55.5	8.3	25.2
Meerkat2-30B-A3B	12.4	57.2	9.3	26.3
Qwen3-32B	13.6	54.4	9.1	25.7
Meerkat2-32B	14.6	51.2	11.2	25.7

Table 2: Performance on additional medical benchmarks. **DiagnosisArena** reports top-5 diagnosis accuracy; **MediQ** reports answer accuracy after interactive questioning; **HealthBench Hard** reports rubric-based scores. All values are on a 0–100 scale.

Cross-Lingual Robustness Meerkat2 also closes the cross-lingual performance gap that typically penalizes non-English queries. On the Doctor subset in **Instant** mode, Qwen3-14B drops by 8.5%p when answering in Korean rather than English, whereas Meerkat2-14B surpasses its English score by 2.1%p. Reducing this cross-lingual gap degradation is important for real-world deployment, since local clinicians predominantly require accurate responses in their native language.

4.3 Beyond Licensing Examinations

Beyond licensing-exam accuracy, we further examine Meerkat2 on three benchmarks that probe broader clinical capabilities. **DiagnosisArena** (Zhu

Model	MedQA	KorMedMCQA		Avg.
		English	Korean	
MedGemma-4B	64.5	55.7	52.2	57.5
Gemma3-4B	52.8	51.2	48.4	50.8
+ Ours	76.0	70.7	69.9	72.2
Phi4-mini	59.7	38.9	33.8	44.1
+ Ours	66.7	61.4	52.6	60.2

Table 3: Backbone transferability. Applying our data mixture to Gemma3-4B and Phi4-mini yields consistent gains on both MedQA and KorMedMCQA. MedGemma-4B, a medically specialized model from the same backbone, is included as a reference.

et al., 2026) evaluates differential diagnosis on real patient case reports curated from top-tier medical journals. **MediQ** (Li et al., 2024b) measures interactive clinical reasoning, where the model is given only minimal initial information and must ask follow-up questions before committing to an answer. **HealthBench Hard** (Arora et al., 2025) contains difficult open-ended health conversations authored by physicians, with responses scored against case-specific rubrics.

Results Table 2 reports the results. Meerkat2 outperforms its Qwen3 counterpart on DiagnosisArena and HealthBench Hard at every model size, and improves the average score at 14B (22.9 to

Model	General Benchmarks							Medical Benchmarks	
	BBH	IFEval	MT-Eval	BFCLv3	TruthfulQA	GSM8K	Avg.	MedSafetyBench	Medical Avg.
Qwen3-14B	67.5	90.5	75.9	76.2	77.6	89.9	79.6	84.4	78.0
Meerkat2-14B	69.9	86.1	67.6	73.3	70.4	89.3	76.1	93.4	85.4
w/o General Mixture	66.5	81.7	59.1	69.8	68.7	88.4	72.4	85.4	85.9

Table 4: Ablation on the general instruction mixture. Training without the general subset causes substantial degradation on general benchmarks, particularly multi-turn conversation (MT-Eval) and instruction following (IFEval), while medical performance remains comparable. The safety subset within the general mixture is responsible for a large portion of the gain on MedSafetyBench. The **Medical Avg.** is computed over MedQA and KorMedMCQA.

27.4) and 30B (25.2 to 26.3) while remaining on par at 32B. These results suggest that the benefits of our training procedure are not confined to exam-style multiple-choice questions but also transfer to differential diagnosis and open-ended clinical conversation, which more closely resemble actual clinical use.

4.4 Backbone Transferability

To verify that our training recipe is not tied to the Qwen3 family, we apply the identical pipeline and data mixture to two distinct backbones, Gemma3-4B (Gemma Team, 2025) and Phi4-mini (Abouelenin et al., 2025), and evaluate on MedQA and KorMedMCQA. We additionally include MedGemma-4B (Sellersgren et al., 2025) as a reference, since it is built on the same Gemma3-4B backbone.

Table 3 reports the results. Our data mixture yields substantial gains on both backbones, improving the average score of Gemma3-4B from 50.8 to 72.2 and that of Phi4-mini from 44.1 to 60.2. Notably, the Gemma3-4B model trained on our mixture outperforms MedGemma-4B, an existing medically specialized model fine-tuned from the same base, by 14.7 points on average. These results indicate that the benefits of our training procedure are not specific to a single model family and transfer to other backbones under the same pipeline.

5 Ablation Study

5.1 General Mixture Replay

To verify the role of the general data described in Section 3.4, we compare our full mixture against a variant trained on the medical subset only. We evaluate on 6 general benchmarks, BIG-Bench Hard (Suzgun et al., 2023), IFEval (Zhou et al., 2023), MT-Eval (Kwan et al., 2024), BFCLv3 (Patil et al., 2025), TruthfulQA (Lin et al., 2022), and GSM8K (Cobbe et al., 2021), along with MedSafetyBench (Han et al., 2024) for medical safety.

As shown in Table 4, removing the general subset causes a substantial drop in general capabil-

Model	Dentist	Doctor	Nurse	Pharm
Qwen3 14B	50.0	57.7	52.9	55.7
+ Distillation	48.3	57.7	50.0	49.2
+ Synthesis (Ours)	56.7	71.2	69.1	57.4
Qwen3 30B A3B	56.7	61.5	54.4	59.0
+ Distillation	50.0	55.8	50.0	50.8
+ Synthesis (Ours)	56.7	73.1	69.1	59.0
Qwen3 32B	53.3	63.5	52.9	60.7
+ Distillation	51.7	57.7	57.4	54.1
+ Synthesis (Ours)	68.3	84.6	77.9	70.5
EXAONE 4.0.1 32B	56.7	59.6	61.8	60.7
Solar Pro 2	55.0	75.0	66.2	70.5
Qwen3 235B A22B	53.3	69.2	61.8	60.7
GPT-5.1	63.3	76.9	76.5	67.2

Table 5: Ablation study on the effectiveness of different data construction methods for the **Law** subsets in KorMedMCQA. **Synthesis**, our proposed method using curated statutes, significantly outperforms baselines.

ity, with the general average falling from 79.6 to 72.4, while the full mixture retains 76.1 and largely preserves the instruction-following and multi-turn behavior. The medical average is nearly identical between the two settings (85.4 vs. 85.9), confirming that the general subset mitigates catastrophic forgetting without sacrificing medical performance. We also observe that MedSafetyBench improves from 85.4 to 93.4 with the full mixture, suggesting that the safety subset contributes to medical safety in addition to its role in mitigating forgetting.

5.2 Law QA Synthesis

To verify the effectiveness of synthesizing law QA pairs explicitly from Korean Medical Statutes, we compare our approach against standard knowledge distillation, which uses rationales from the teacher model. As presented in Table 5, distillation alone yields negligible gains on the legal subsets. In contrast, QA synthesis results in substantial improvements, surpassing existing powerful models. To assess the factual quality of the two data sources, we apply the LF-Eval framework (Lee et al., 2025), which measures legal fidelity against the underlying statutes. Synthesized QA scores 98.7, far above

Model	MedQA	KorMedMCQA	
		English	Korean
Qwen3-14B	76.0 (492.2)	75.5 (435.6)	72.4 (567.4)
+ Verbose CoTs	85.4 (669.3)	82.9 (693.5)	80.6 (965.5)
+ Concise CoTs	83.9 (299.9)	82.0 (236.7)	81.1 (350.8)
+ VDD (Ours)	85.2 (309.9)	82.5 (243.6)	82.1 (358.3)
Qwen3-235B-A22B	89.8 (340.4)	86.7 (302.5)	88.0 (606.5)
GPT-5.1	93.2 (211.5)	91.0 (154.1)	90.8 (234.4)

Table 6: Ablation study on reasoning efficiency. Values represent **Accuracy (Avg. Tokens)**. "VDD" refers to Verbosity-Decoupled Distillation. It achieves the best balance of high accuracy and low token cost.

the 79.0 obtained by distilled rationales, indicating that teacher-generated rationales frequently introduce factual or legal inaccuracies. These findings suggest that for knowledge-intensive domains requiring precise memorization, such as law, deriving structured rationales directly from original documents is more effective than relying on a teacher model’s inaccurate and unreliable outputs.

5.3 Verbosity-Decoupled Distillation

To examine the effect of Verbosity-Decoupled Distillation, we compare it against two single-style baselines, distillation from *Verbose CoTs* only and from *Concise CoTs* only. Table 6 shows the expected trade-off between the two. On MedQA, verbose distillation yields 85.4% accuracy while concise distillation yields only 83.9%, but the verbose variant more than doubles the average generation length (669.3 vs. 299.9 tokens). VDD resolves this trade-off rather than splitting it. The student attains 85.2% accuracy, slightly below the verbose baseline, while generating only 309.9 tokens on average, comparable to the concise variant and a 54% reduction relative to verbose distillation. The same pattern holds on the English and Korean subsets of KorMedMCQA, where VDD reduces the average token count by 65% and 63% of the verbose baseline without sacrificing accuracy. These results indicate that the verbosity attached to detailed CoTs can be cleanly separated from the reasoning content itself, allowing the student to inherit the reasoning benefits without internalizing the verbose response style. Figure 1 situates this trade-off against other models and shows that Meerkat2 occupies a favorable position on the accuracy-efficiency frontier. A qualitative example is provided in Figure 2.

6 Mechanistic Analysis of VDD

In Section 5.3, we showed that attaching a verbosity instruction to long CoTs prevents the student from

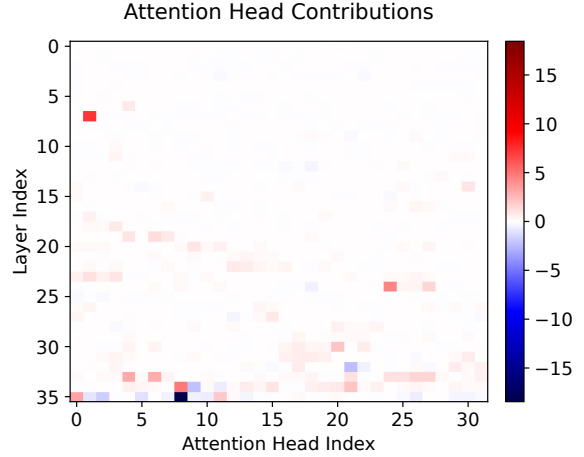


Figure 3: Per-head contributions to the verbose direction across all layers and heads of Qwen3-4B. Red cells write along the verbose direction and blue cells write against it. Verbose generation is driven by a sparse set of attention heads (e.g., L7H1, L24H24, and L33H4).

internalizing verbose generation as its default style. The recipe is simple and the gains are consistent, but it remains unclear why a prompt-level signal is enough to block a behavior that is otherwise easily transferred from teacher to student.

Related strategies have been explored in several lines of work. Kang et al. (2025) train the student on long reasoning traces alongside compressed counterparts under a length condition, recovering reasoning gains on math and common-sense tasks. Chen et al. (2025b) fine-tune with a target persona’s steering vector activated, preventing that trait from being absorbed into the base model. Zhou et al. (2024) pre-train a LoRA (Hu et al., 2022) on harmful data and keep it active during downstream fine-tuning, so that the base model does not update toward the harmful behavior.

What unites these methods is that the undesired behavior is already turned on during training. The intuition is that if the behavior is already present in the model’s output, the gradient has little reason to push the base parameters further in that direction, leaving the behavior tied to its trigger rather than internalized as default. This preventative approach is empirically effective, yet how this conditioning acts on the model internally is rarely analyzed. In this section, we analyze how the verbosity instruction reshapes the gradient updates during distillation, providing gradient-level evidence for how VDD reduces verbosity internalization.

6.1 Finding Verbose Attention Heads

Mechanistic interpretability has shown that distinct behaviors of LLMs are often carried by specific

attention heads, such as Olsson et al. (2022), Li et al. (2023), and Wu et al. (2025). As a first step toward understanding how VDD shapes the model, we figure out the heads that drive verbose generation, so that subsequent analysis can examine how the training reinforces these heads.

To identify such heads, we adopt the procedure of Sun et al. (2025). Given a set of prompts paired with a long and a short response, we first extract a verbose direction at each layer as the difference between the mean attention activations over the response tokens of the two response types:

$$v_l^{\text{attn}} = \bar{r}_{l,\text{long}}^{\text{attn}} - \bar{r}_{l,\text{short}}^{\text{attn}}, \quad \hat{v}_l^{\text{attn}} = \frac{v_l^{\text{attn}}}{\|v_l^{\text{attn}}\|}, \quad (1)$$

where $\bar{r}_{l,\text{long}}^{\text{attn}}$ and $\bar{r}_{l,\text{short}}^{\text{attn}}$ denote the mean attention activation at layer l over the response tokens of the long and short variants, respectively. We then compute the contribution of each attention head h at layer l to this verbose direction:

$$\bar{C}_{l,\text{long}}^h = \langle \bar{A}_l^h W_{l,o}^h, \hat{v}_l^{\text{attn}} \rangle, \quad (2)$$

where $\bar{A}_l^h$ is the mean per-head activation before the output projection and $W_{l,o}^h$ is the corresponding weight. A large positive contribution leads to the residual stream along the verbose direction.

We apply this analysis to Qwen3-4B on a 2.6K subset of our training data, using the verbose and concise pairs from the same prompts. Figure 3 shows the per-head contributions. A small number of heads stand out, including L7H1, L24H24, L33H4, L34H8, and L35H0 (where L l H h denotes head h at layer l), confirming that verbose generation is driven by a sparse set of attention heads.

6.2 Gradients to the Verbose Direction

We now quantify how each training setting affects the verbose heads identified in Section 6.1. Inspired by Kramár et al. (2024), we derive a first-order approximation of the change in verbose contribution, $\Delta \bar{C}_{l,\text{long}}^h$, directly from the gradient of the loss. For a parameter update with learning rate η and cross-entropy loss $\mathcal{L}$ over the response tokens,

$$W_{l,o}^h \leftarrow W_{l,o}^h - \eta \frac{\partial \mathcal{L}}{\partial W_{l,o}^h}, \quad (3)$$

$$\Delta W_{l,o}^h = -\eta \frac{\partial \mathcal{L}}{\partial W_{l,o}^h} = -\eta (\bar{A}_l^h)^T \frac{\partial \mathcal{L}}{\partial \bar{O}_l^h}, \quad (4)$$

where $\bar{O}_l^h = \bar{A}_l^h W_{l,o}^h$. By Equation 2, we have

$$\frac{\partial \bar{C}_{l,\text{long}}^h}{\partial W_{l,o}^h} = (\bar{A}_l^h)^T (\hat{v}_l^{\text{attn}})^T. \quad (5)$$

Dataset	Top 1%	Top 10%	All Heads
Concise CoTs	-12.2	-30.4	-63.0
Verbose CoTs	31.1	89.7	251.2
w/ System Prompt	-7.9	-19.0	-37.1

Table 7: Aggregate change $\Delta \bar{C}$ in verbose-direction contribution, summed over the top 1%, top 10%, and all attention heads. VDD reverses the sign of the update, matching the pattern of concise distillation.

A first-order Taylor approximation then gives

$$\begin{aligned} \Delta \bar{C}_{l,\text{long}}^h &\approx \left\langle \frac{\partial \bar{C}_{l,\text{long}}^h}{\partial W_{l,o}^h}, \Delta W_{l,o}^h \right\rangle \\ &= -\eta \left\langle (\bar{A}_l^h)^T (\hat{v}_l^{\text{attn}})^T, (\bar{A}_l^h)^T \frac{\partial \mathcal{L}}{\partial \bar{O}_l^h} \right\rangle \\ &= -\eta \|\bar{A}_l^h\|^2 \left(\frac{\partial \mathcal{L}}{\partial \bar{O}_l^h} \cdot \hat{v}_l^{\text{attn}} \right). \end{aligned} \quad (6)$$

We compute it under three training settings: (1) Concise CoTs, (2) Verbose CoTs, and (3) Verbose CoTs paired with a verbosity instruction which corresponds to VDD. Activations $\bar{A}_l^h$ and gradients $\partial \mathcal{L} / \partial \bar{O}_l^h$ are collected from a single backward pass on a subset of 100 examples, and $\Delta \bar{C}_{l,\text{long}}^h$ is aggregated over the top 1%, top 10%, and all attention heads ranked by verbose contribution.

Table 7 reports the result. Verbose CoTs alone produce a strong positive update along the verbose direction (+251.2), while concise distillation yields a negative update (−63.0) that reduces response length. Attaching a verbosity instruction to the verbose CoTs reverses the sign to −37.1, closely matching the concise setting, and the same pattern holds at the top 1% and top 10% heads. This confirms the preventative training intuition. Activating the undesired behavior through the prompt cancels the gradient that would otherwise push the parameters toward it.

7 Conclusion

We present **Meerkat2**, a family of open medical models trained with knowledge-matched supervision: clinical reasoning distilled from strong teachers, and QA pairs synthesized from Korean medical statutes. We introduce VDD, which retains the reasoning gains of long rationales without internalizing verbose output. Meerkat2 improves over comparable models on MedQA and KorMedMCQA and transfers to other backbones, and our analysis shows that the verbosity instruction cancels the gradient toward verbose generation, providing a principled account for preventative training.

Acknowledgement

This research was supported by "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea [Ministry of Science and ICT], the Ministry of Science and ICT (MSIT) and National IT Industry Promotion Agency (NIPA) through the Hyperscale AI Healthcare Service Support Program [H0701-25-1003], a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [RS-2025-25462758, RS-2022-KH129295, HR20C002103], the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [RS-2026-25510253, NRF-2023R1A2C3004176, RS-2023-00262002], (MSIT and MOE) [RS-2025-16652968].

Limitations

Our training pipeline relies on large external teacher models (DeepSeek-V4-Pro and Qwen3-235B-A22B) for synthesizing clinical reasoning trajectories and biomedical question-answer pairs. The synthesis stage incurs nontrivial inference cost. Our human evaluation is similarly conducted at a limited scale, since recruiting clinicians for large-scale comparative assessment is costly. Broader physician studies covering more cases and a wider panel of specialists would be needed to fully characterize the model’s clinical behavior.

Our mechanistic analysis of verbose attention heads is conducted on a single backbone (Qwen3-4B). The mechanism we identify is expected to extend to other transformer architectures, though a systematic verification across additional model families is beyond the scope of this study. Meerkat2 is also adapted to Korean clinical contexts, and applying the same pipeline to other languages or healthcare systems would require constructing analogous jurisdiction-specific resources, including legal corpora and clinical benchmarks.

Ethics Statement

Our training data is drawn from publicly available benchmarks, instruction-tuning datasets, and Korean statutes accessible through public legal information services. We do not use any identifiable patient records. Synthetic QA pairs and rationales are generated by DeepSeek-V4-Pro and Qwen3-235B-A22B, and we retain only those whose final

answers match the reference, although we acknowledge that generated content may still contain inaccuracies.

Meerkat2 is intended for research and educational use. Its outputs are not a substitute for professional medical diagnosis or treatment, and clinical decisions should not be made on the basis of model responses without expert verification. To address safety, our training mixture includes a refusal subset designed to reject harmful prompts, and we report the resulting behavior on MedSafetyBench.

References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025a. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.
- Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025b. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. *MEDITRON-70B: Scaling medical pretraining for large language models*. *Preprint*, arXiv:2311.16079.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

- DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.
- Gemma Team. 2025. [Gemma 3](#).
- Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. [OpenThoughts: Data recipes for reasoning models](#). *Preprint*, arXiv:2506.04178.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and 1 others. 2025. [DeepSeek-R1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024. [Medsafetybench: Evaluating and improving the medical safety of large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 33423–33454. Curran Associates, Inc.
- Healthcare AI Research Institute(HARI) of Seoul National University Hospital(SNUH). 2025. [hari-q3-14b](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Hugging Face. 2025a. [Open R1: A fully open reproduction of DeepSeek-R1](#).
- Hugging Face. 2025b. [SmolTalk2](#).
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#). *Applied Sciences*, 11(14).
- Yu Kang, Xianghui Sun, Liangyu Chen, and Wei Zou. 2025. [C3ot: Generating shorter chain-of-thought without compromising effectiveness](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24312–24320.
- Kensaku Kawamoto, Caitlin A. Houlihan, E. Andrew Balas, and David F. Lobach. 2005. [Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success](#). *BMJ*, 330(7494):765.
- Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. 2025. [Small language models learn enhanced reasoning skills from medical textbooks](#). *npj Digital Medicine*, 8(1):240.
- Yoon Kim and Alexander M. Rush. 2016. [Sequence-level knowledge distillation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- János Kramár, Tom Lieberum, Rohin Shah, and Neel Nanda. 2024. [Atp*: An efficient and scalable method for localizing llm behaviour to components](#). *arXiv preprint arXiv:2403.00745*.
- Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. [MT-eval: A multi-turn capabilities evaluation benchmark for large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.
- Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. 2024. [KorMedMCQA: Multi-choice question answering benchmark for korean healthcare professional licensing examinations](#). *Preprint*, arXiv:2403.01469.
- Maxime Labonne. 2024. [Fine-tune llama 3.1 ultra-efficiently with unsloth](#).
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xixi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, and 4 others. 2025. [Tulu 3: Pushing frontiers in open language model post-training](#). In *Second Conference on Language Modeling*.
- Jihyung Lee, Daehui Kim, Seonjeong Hwang, Hyounghun Kim, and Gary Lee. 2025. [KoBLEX: Open legal question answering with multi-hop reasoning](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4019–4053, Suzhou, China. Association for Computational Linguistics.
- LG AI Research. 2025. [EXAONE 4.0: Unified large language models integrating non-reasoning and reasoning modes](#). *arXiv preprint arXiv:2507.11407*.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat,

- Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others. 2024a. [Datacomp-lm: In search of the next generation of training sets for language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 14200–14282. Curran Associates, Inc.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 41451–41530. Curran Associates, Inc.
- Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024b. [Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888. Curran Associates, Inc.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [TruthfulQA: Measuring how models mimic human falsehoods](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.
- ModelScope Team. 2024. [EvalScope: Evaluation framework for large models](#).
- Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. 2025. [AIMO-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset](#). *Preprint*, arXiv:2504.16891.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*. <https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html>.
- OpenAI. 2024. [OpenAI o1 system card](#). *Preprint*, arXiv:2412.16720.
- OpenAI. 2025. [GPT-5.1 instant and GPT-5.1 thinking system card addendum](#).
- Shishir G Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. 2025. [The berkeley function calling leaderboard \(BFCL\): From tool use to agentic evaluation of large language models](#). In *Forty-second International Conference on Machine Learning*.
- Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. [Towards building multilingual language model for medicine](#). *Nature Communications*, 15(1):8384.
- Zhongxi Qiu, Zhang Zhang, Yan Hu, Heng Li, and Jiang Liu. 2025. [Open-Medical-R1: How to choose data for rlvr training at medicine domain](#). *Preprint*, arXiv:2504.13950.
- Qwen Team. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qwen Team. 2025b. [QwQ-32B: Embracing the power of reinforcement learning](#).
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. [Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, page 3505–3506, New York, NY, USA. Association for Computing Machinery.
- David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019. [Experience replay for continual learning](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. [Fine-tuned language models are continual learners](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. [MedGemma technical report](#). *Preprint*, arXiv:2507.05201.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. [KMMLU: Measuring massive multitask language understanding in Korean](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.

- Chung-En Sun, Ge Yan, and Tsui-Wei Weng. 2025. [ThinkEdit: Interpretable weight editing to mitigate overly short thinking in reasoning models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17001–17025, Suzhou, China. Association for Computational Linguistics.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging BIG-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. [Large language models in medicine](#). *Nature Medicine*, 29(8):1930–1940.
- Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2024. [Opportunities and challenges for chatgpt and large language models in biomedicine and health](#). *Briefings in Bioinformatics*, 25(1):bbad493.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Upstage AI. 2025. [Solar Pro 2](#).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn, and Jakob Nikolas Kather. 2024. [Privacy-preserving large language models for structured medical information retrieval](#). *npj Digital Medicine*, 7(1):257.
- Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2025. [Retrieval head mechanistically explains long-context factuality](#). In *International Conference on Learning Representations*, volume 2025, pages 62143–62156.
- Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason Weston, and Xian Li. 2025. [Naturalreasoning: Reasoning in the wild with 2.8m challenging questions](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhao Chen. 2024. [MAMmoTH2: Scaling instructions from the web](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 90629–90660. Curran Associates, Inc.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. [Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model?](#) In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Daniel Shao, Yong Hoe Koo, Ko Minhyeok, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. 2025. [Med-PRM: Medical reasoning models with stepwise, guideline-verified process rewards](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16554–16571, Suzhou, China. Association for Computational Linguistics.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. [HuatuoGPT, towards taming language model to be a doctor](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.
- Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025. [Med-RLVR: Emerging medical reasoning from a 3b base model via reinforcement learning](#). *Preprint*, arXiv:2502.19655.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.
- Xin Zhou, Yi Lu, Ruotian Ma, Yujian Wei, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. [Making harmful behaviors unlearnable for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10258–10273, Bangkok, Thailand. Association for Computational Linguistics.
- Yakun Zhu, Zhongzhen Huang, Linjie Mu, Yutong Huang, Wei Nie, Jiaji Liu, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2026. [DiagnosisArena: Benchmarking diagnostic reasoning for large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 3074–3098, San Diego, California, United States. Association for Computational Linguistics.

A Appendix

A.1 Supplementary Details on Data Curation

To generate clinical reasoning data, we employ 6 distinct prompt templates, shown in Figure 4, designed to elicit responses in either English or Korean. As illustrated in the bottom of the figure, the prompt used for evaluation differs from the prompts used during training. We implement a rejection sampling strategy, retaining only the rationales that correctly answered the question. We allow up to 10 generation attempts per question to obtain a valid response. To enhance data diversity, we generate multiple rationale variations for each dataset: 4 for MedQA, 8 for KorMedMCQA, and 1 for MedBooks-18. Finally, excessively long responses are excluded to maintain data quality.

For law synthesis, we identify 12 statutes governing medicine and healthcare within the KorMedMCQA benchmark, which are summarized in Table 8. As described in Section 3.2, we utilize the RecursiveCharacterTextSplitter² with a chunk size of 3,000 characters and an overlap of 1,000 characters to mitigate information loss due to truncation. The instruction used for synthesizing QA pairs is presented in Figure 6 (see Figure 7 for the English translation).

A.1.1 General LLM Capabilities

To ensure the model retains general capabilities beyond biomedical knowledge, we incorporate a subset of general instruction-tuning datasets from SmolTalk2 and Tulu-3. The detailed mixture configuration is listed in Table 12.

System Prompt To preserve the model’s ability to adhere to system prompts, including role-playing, specific response styles, and situational constraints, we include smoltalk_systemchats_Qwen3_32B_think and smoltalk_smollm3_systemchats_30k_no_think from the SmolTalk2 dataset.

Multilingual To support multilingual capabilities, we select aya_dataset_Qwen3_32B_think, smoltalk_multilingual8_Qwen3_32B_think, and smoltalk_multilingual_8languages_lang_5_no_think from SmolTalk2. Furthermore, we utilize a Korean-translated version of FineTome (Labonne, 2024)³ with Qwen3 32B Instant

²https://reference.langchain.com/python/langchain_text_splitters

³<https://huggingface.co/datasets/lemon-mint/Korean-FineTome-100k>

Korean Statute Name	English Title
의료법	Medical Service Act
약사법	Pharmaceutical Affairs Act
지역보건법	Regional Public Health Act
국민건강보험법	National Health Insurance Act
마약류 관리에 관한 법률	Narcotics Control Act
보건의료기본법	Framework Act on Health and Medical Services
국민건강증진법	National Health Promotion Act
감염병의 예방 및 관리에 관한 법률	Infectious Disease Control and Prevention Act
검역법	Quarantine Act
후천성면역결핍증 예방법	Prevention of Acquired Immunodeficiency Syndrome Act
응급의료에 관한 법률	Emergency Medical Service Act
혈액관리법	Blood Management Act

Table 8: List of major Korean medical and public health statutes with their official English translations.

and Think annotations. In total, the dataset covers 22 languages, aiming to preserve multilingual knowledge with a minimal data footprint.

Multi-Turn Conversation To enhance the multi-turn dialogue capabilities, we include smoltalk_everyday_convs_reasoning_Qwen3_32B_think, smoltalk_smollm3_everyday_conversations_no_think, and smoltalk_smollm3_smol_magpie_ultra_no_think from SmolTalk2, which contain conversations consisting of average 6 turns.

Instruction Following For precise instruction following, we utilize smoltalk_smollm3_explore_instruct_rewriting_no_think and smoltalk_smollm3_smol_summarize_no_think from SmolTalk2, alongside tulu_3_sft_personas_instruction_following_no_think from Tulu-3, which focuses on strict adherence to user constraints.

Safety Finally, to ensure safety and robust refusal of harmful queries, we employ tulu_v3.9_wildjailbreak_decontaminated_50k and tulu_v3.9_synthetic_finalresp_wildguardmixtrain_decontaminated_50k.

A.2 Decontamination

We check for potential overlap between our training data and the test sets of MedQA and KorMedMCQA by searching for shared 10-grams, following the criterion used in prior work (Touvron et al., 2023; Li et al., 2024a). We confirm that no 10-gram overlap is found between any training and

Model	English Response	Korean Response
Qwen3-14B	71.0	63.2
Meerkat2-14B	83.9	88.4
Qwen3-30B-A3B	74.2	70.3
Meerkat2-30B-A3B	83.2	83.2
Qwen3-32B	81.3	73.6
Meerkat2-32B	89.7	89.0
Qwen3-235B-A22B	85.2	88.4
GPT-5.1	94.2	92.9

Table 9: Performance on the 89th Korean Medical Licensing Examination (KMLE 2025).

test examples, indicating that our reported evaluation results are not affected by data leakage.

A.3 Training Details

We train Meerkat2 using 8 NVIDIA A100 GPUs. The training configuration utilizes a maximum sequence length of 8192 tokens, a learning rate of $5e-5$, and a global batch size of 256. We use AdamW with DeepSpeed Zero3 (Rasley et al., 2020). All models are fine-tuned for a single epoch. To enforce robust reasoning capabilities, we apply loss masking to both the user prompts and the empty thought blocks. This strategy makes the model to initiate reasoning by default, ensuring it generates a reasoning content unless it is explicitly disabled.

A.4 Performance on KMLE 2025

Because our models were trained on KorMedMCQA, which cover the period from 2012 to 2024, it is essential to assess their generalizability to unseen data beyond this timeframe. To this end, we collected the 89th Korean Medical Licensing Examination (KMLE 2025) from the Korea Health Personnel Licensing Examination Institute⁴.

We extracted questions and answers from the official PDF documents using OCR. To maintain consistency with the KorMedMCQA format, we excluded multimodal or restricted questions. This curation process allows us to evaluate the models on general medical knowledge and clinical reasoning tasks that post-date the 2024 knowledge cutoff.

The results are presented in Table 9. Meerkat2 consistently outperforms the original Qwen3 models, and the 32B variant surpasses even the significantly larger Qwen3-235B-A22B. Furthermore,

⁴<https://www.kuksiwon.or.kr/>

Model	MedQA	KorMedMCQA	
		English	Korean
Qwen3-14B	76.5 $\pm$ 0.7 (514.4 $\pm$ 2.4)	76.2 $\pm$ 0.2 (439.0 $\pm$ 1.4)	71.9 $\pm$ 0.5 (571.1 $\pm$ 0.8)
+ Verbose CoTs	84.8 $\pm$ 0.6 (684.1 $\pm$ 4.9)	83.7 $\pm$ 0.2 (735.8 $\pm$ 2.7)	82.4 $\pm$ 0.4 (1007.2 $\pm$ 7.8)
+ Concise CoTs	83.8 $\pm$ 0.3 (317.1 $\pm$ 2.3)	82.2 $\pm$ 0.3 (244.7 $\pm$ 1.5)	80.6 $\pm$ 0.4 (364.2 $\pm$ 1.7)
+ VDD (Ours)	84.7 $\pm$ 0.4 (321.2 $\pm$ 1.8)	83.0 $\pm$ 0.2 (246.6 $\pm$ 0.9)	82.4 $\pm$ 0.4 (367.1 $\pm$ 4.4)

Table 10: Ablation study on the robustness of VDD across random seeds. Values represent Accuracy (Avg. Tokens), reported as mean $\pm$ standard deviation over three random seeds under random sampling. VDD remains the best balance of high accuracy and low cost.

Model	Simple	EvalScope	Meerkat1
Qwen3-14B	76.7 (290.6)	76.0 (492.2)	76.2 (559.2)
+ Verbose CoTs	85.0 (683.6)	85.4 (669.3)	85.6 (680.1)
+ Concise CoTs	83.4 (339.2)	83.9 (299.9)	83.0 (286.1)
+ VDD (Ours)	85.0 (289.7)	85.2 (309.9)	85.3 (347.5)

Table 11: Ablation study on the robustness of VDD across system prompts on MedQA. System Prompt 1 (Simple), 2 (EvalScope), and 3 (Meerkat1) are detailed in Figure 5. Values represent Accuracy (Avg. Tokens). VDD remains the best balance of high accuracy and low token cost across all three prompts.

while the baseline models exhibit a notable performance disparity depending on the response language (with a gap of up to 7.8%p), Meerkat2 successfully narrows this gap, demonstrating robust multilingual consistency.

A.5 Robustness of VDD

To examine the robustness of VDD, we vary two factors at evaluation time: the random seed used for decoding and the system prompt presented to the model.

A.5.1 Variation Across Decoding Seeds

We re-evaluate each benchmark with three random seeds under random sampling, and report accuracy and average token count as mean $\pm$ stdev in Table 10. The standard deviations are small relative to the gaps between methods, and the ranking reported in Table 6 holds under all three seeds.

A.5.2 Variation Across System Prompts

We additionally evaluate the models on MedQA under three system prompts, shown in Figure 5. Across all three prompts, VDD matches Verbose CoTs in accuracy while using substantially fewer tokens (Table 11).

System Prompt A1: $\LaTeX$Boxed Format ($\boxed{}$) Answer the following multiple-choice question. Think step by step and infer the correct answer. Respond only in English. At the end of your response, output the answer enclosed in "$\boxed{A/B/C/D/E}$". 다음의 주어진 객관식 질문에 답하십시오. 순차적으로 생각한 후 정답을 추론하세요. 한국어로만 응답해야 합니다. 응답 마지막에 정답을 "$\boxed{A/B/C/D/E}$"로 감싸서 출력하세요.
System Prompt A2: Natural Language Suffix Respond only in English. Answer the multiple-choice question step by step. At the end, write: "The answer is (A/B/C/D/E)." 객관식 문제를 단계적으로 풀고 한국어로만 응답하세요. 마지막에는 이렇게 적으세요: "정답은 (A/B/C/D/E)이다."
System Prompt A3: Explicit Label Format Think step by step and then choose the correct option. Use English only. Conclude with: "Answer: A/B/C/D/E" 단계적으로 사고한 뒤 정답을 고르시오. 한국어로만 응답할 것. 마지막에 "정답: A/B/C/D/E"를 적으시오.
System Prompt A4: JSON Structured Output Answer the following multiple-choice question step by step. Respond only in English. End with: <pre>```json {"answer": "A/B/C/D/E"}</pre> 다음 객관식 문제에 단계적으로 답하세요. 한국어로만 응답하세요. 마지막은 다음과 같이 마무리하세요: <pre>```json {"answer": "A/B/C/D/E"}</pre>
System Prompt A5: Evaluation Standard (EvalScope) Answer the following multiple choice question. The last line of your response should be of the following format: 'ANSWER: \$LETTER' (without quotes) where LETTER is one of A/B/C/D/E. Think step by step before answering. Answer only in English. 다음 객관식 문제에 답하세요. 응답의 마지막 줄은 따옴표 없이 'ANSWER: \$LETTER' 형식이어야 하며, LETTER는 A/B/C/D/E 중 하나입니다. 답변하기 전에 단계별로 생각하세요. 한국어로 응답하세요.

Figure 4: System prompts used for training and evaluation. System Prompts A1–A4 employ diverse formatting instructions ($\LaTeX$, Natural Language, Label, JSON) to generate heterogeneous training samples and prevent overfitting. System Prompt A5 represents the fixed, standard format used during the evaluation phase.

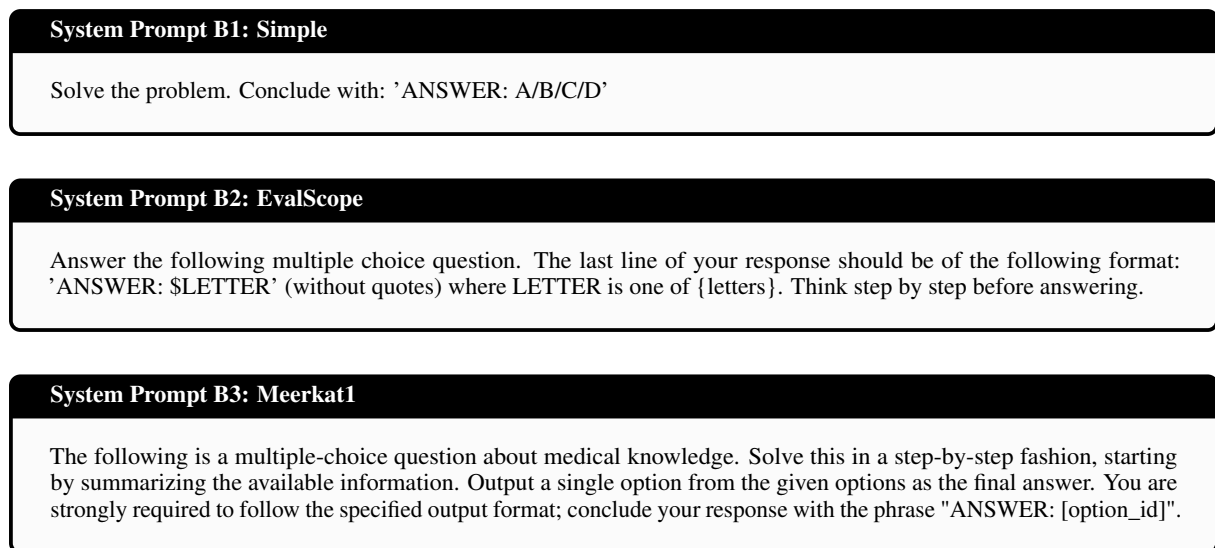


Figure 5: System prompts used to evaluate the robustness of VDD to prompt formulation (Section A.5). System Prompt B1 is a minimal instruction, System Prompt B2 follows the EvalScope evaluation standard, and System Prompt B3 adopts a Meerkat1-style medical instruction.

Instruction for Law QA Synthesis

다음은 대한민국 「{name}」 중 일부를 발췌한 내용입니다. 주어진 예시 문제를 참고하여, 해당 법률 자료에 대한 독립적인 객관식 문제를 생성하세요. 지침은 다음과 같습니다:

1. 주어진 법률 자료에 한하여 문제를 생성해야 합니다. 존재하지 않는 내용이나 추측성 정보는 반드시 없어야 합니다.
2. 잘린 문항이나 전후 내용을 정확히 파악하기 힘든 경우는 문제를 만들기보다는 무시하세요. 만일 주어진 법률 자료가 문제를 생성하는 데에 적합하지 않다면, 문제를 생성하지 않아도 됩니다. 불필요한 정보를 가지고 굳이 문제를 생성하려 하지 마세요.
3. 주어진 발췌문이 실제 문제 풀이 시에는 주어지지 않습니다. 모든 문제는 독립적이어야 합니다. 응시자가 문제 내에서 모든 정보 혹은 맥락을 파악할 수 있도록 하세요. 즉, 문제와 풀이 모두 **발췌문이 원래부터 존재하지 않는 것처럼** 작성되어야 합니다. 발췌문은 문제를 생성하기 위함임을 명심하세요.
4. 하단의 예시를 참고하여, 최대 8개의 질문을 하고 올바른 답변을 제시하세요. 순서는 무작위여도 좋습니다. 학습에 도움이 되도록 근거를 포함하여 자세한 설명과 같이 정답 예시를 포함해야 합니다. 구조는 다음과 같이 작성하시오:

[QUESTION 1]

(객관식 질문)

[SOLUTION 1]

(암기에 도움이 되는 지식을 포함하여 500단어 이상의 설명)

[ANSWER 1]

(ONE OF A/B/C/D/E)

[QUESTION 2]

...

5. 단순히 법률 근거를 묻는 질문 외에, 사소한 추론이 필요한 문제를 희소하게 포함해도 좋습니다. 문제가 너무 어렵거나 복잡해서는 안됩니다. 문제 풀이 설명은 추론 과정이 정확해야 합니다.
6. 각 항목에는 일반 텍스트를 사용하세요. Markdown 등의 특수 포맷은 허용하지 않습니다.

<chunk_of_law>

{law_chunk}

</chunk_of_law>

<example_questions>

{examples}

</example_questions>

Figure 6: Instruction prompt used for synthesizing law QA pairs from Korean medical statutes.

Instruction for Law QA Synthesis (English Translation)

The following is an excerpt from the Republic of Korea's 「**{name}**」. Referring to the provided example questions, create independent multiple-choice questions based purely on the relevant legal material. The guidelines are as follows:

1. Questions must be generated solely based on the given legal material. Non-existent content or speculative information must absolutely not be included.
2. If an item is cut off or the preceding/following context is difficult to grasp accurately, ignore it rather than creating a question. If the given legal material is not suitable for generating questions, you do not have to generate questions. Do not force the creation of questions using unnecessary information.
3. The provided excerpt will not be given during actual problem-solving. All questions must be independent. Ensure the examinee can grasp all information or context within the question itself. In other words, both the question and the solution must be written **as if the excerpt did not exist originally**. Keep in mind that the excerpt is only for the purpose of generating the questions.
4. Referencing the examples below, ask a maximum of 8 questions and present the correct answers. The order may be random. Include detailed explanations with rationale to aid learning. The structure should be written as follows:

[QUESTION 1]
(Multiple choice question)

[SOLUTION 1]
(Explanation of 500+ words including knowledge helpful for memorization)

[ANSWER 1]
(ONE OF A/B/C/D/E)

[QUESTION 2]
...

5. In addition to questions simply asking for legal basis, you may rarely include questions requiring minor reasoning. The questions must not be too difficult or complex. The reasoning process in the explanation must be accurate.
6. Use plain text for each item. Special formats like Markdown are not permitted.

<chunk_of_law>
{law_chunk}
</chunk_of_law>

<example_questions>
{examples}
</example_questions>

Figure 7: English translation of the instruction prompt used for synthesizing law QA pairs.

Category	Dataset Component	Total	Sampled
Medical Domain			
Medical Law	korean_medical_laws_qa	18,721	18,721
Medical MCQA Reasoning Distillation	kormedmcqa_dsv4pro_concise	67,456	67,456
	kormedmcqa_dsv4pro_verbose	67,631	67,631
	kormedmcqa_qwen3_235b_a22b_thinking	66,326	66,326
	medbooks18_dsv4pro_concise	77,120	77,120
	medbooks18_dsv4pro_verbose	76,993	76,993
	medqa_4opt_dsv4pro_concise	79,391	79,391
	medqa_4opt_dsv4pro_verbose	79,755	79,755
	medqa_4opt_qwen3_235b_a22b_thinking	78,357	78,357
Open Medical QA	natural_reasoning_medical_subset	19,980	19,979
	openthoughts_medical_subset	1,031	1,031
	webinstruct_medical_subset	36,018	36,018
<i>Medical Subtotal</i>		—	668,778
General Domain			
System Prompt	smoltalk_smollm3_systemchats_30k_no_think	54,217	20,000
	smoltalk_systemchats_Qwen3_32B_think	27,396	19,995
Multilingual	aya_dataset_Qwen3_32B_think	15,204	4,999
	korean_finetome_100k_qwen3_32b_no_think	99,998	20,000
	korean_finetome_100k_qwen3_32b_think	97,433	19,766
	smoltalk_multilingual18_Qwen3_32B_think	244,516	4,959
	smoltalk_multilingual_8languages_lang_5_no_think	254,047	5,000
Multi-Turn Conversation	smoltalk_everyday_convs_reasoning_Qwen3_32B_think	2,056	2,056
	smoltalk_smollm3_everyday_conversations_no_think	2,004	2,004
	smoltalk_smollm3_smol_magpie_ultra_no_think	813,686	20,000
Instruction Following	smoltalk_smollm3_explore_instruct_rewriting_no_think	30,391	10,000
	smoltalk_smollm3_smol_summarize_no_think	96,061	10,000
	tulu_3_sft_personas_instruction_following_no_think	29,970	9,998
Safety	tulu_v3.9_synthetic_finalresp_wildguardmixtrain_decontaminated_50k	50,000	10,000
	tulu_v3.9_wildjailbreak_decontaminated_50k	50,000	10,000
<i>General Subtotal</i>		—	168,777
Total Mixture		—	837,555

Table 12: Composition of the dataset mixture. To preserve general capabilities, we incorporated general instruction-tuning datasets alongside biomedical-specific data.