ControlMed: Adding Reasoning Control to Medical Language Model

Sung-Min Lee **Siyoon Lee** **Juyeon Kim** **Kyoungmin Roh**
Agentic AI Lab, KT
{lee.sungmin, siyoon.lee, jy95.kim, kyoungmin.roh}@kt.com

Abstract

Reasoning Large Language Models (LLMs) with enhanced accuracy and explainability are increasingly being adopted in the medical domain, as the life-critical nature of clinical decision-making demands reliable support. Despite these advancements, existing reasoning LLMs often generate unnecessarily lengthy reasoning processes, leading to significant computational overhead and response latency. These limitations hinder their practical deployment in real-world clinical environments. To address these challenges, we introduce **ControlMed**, a medical language model that enables users to actively control the length of the reasoning process at inference time through fine-grained control markers. ControlMed is trained through a three-stage pipeline: 1) pre-training on a large-scale synthetic medical instruction dataset covering both *direct* and *reasoning responses*; 2) supervised fine-tuning with multi-length reasoning data and explicit length-control markers; and 3) reinforcement learning with model-based reward signals to enhance factual accuracy and response quality. Experimental results on a variety of English and Korean medical benchmarks demonstrate that our model achieves similar or better performance compared to state-of-the-art models. Furthermore, users can flexibly balance reasoning accuracy and computational efficiency by controlling the reasoning length as needed. These findings demonstrate that ControlMed is a practical and adaptable solution for clinical question answering and medical information analysis.

1 Introduction

Large language models (LLMs) have achieved outstanding results in a wide range of medical natural language processing tasks, such as medical question answering, clinical diagnosis, and patient consultation (Su et al., 2025). Their capability to synthesize complex medical knowledge and provide detailed explanations makes them promising tools in both clinical practice and medical education (Li et al., 2024; Singhal et al., 2025). Recent advances have led to the emergence of *medical reasoning models*, such as HuatuoGPT-o1 (Chen et al., 2024a), which further enhance accuracy and reliability by explicitly generating a reasoning process before providing a response.

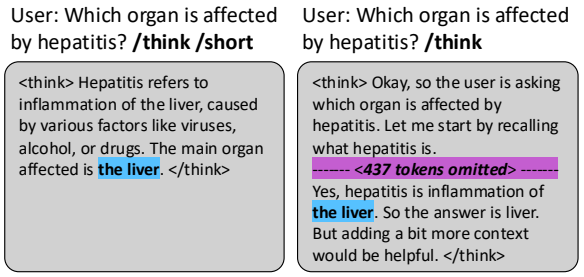

Figure 1: An illustration of the effect of different reasoning control markers on model behavior. When prompted with the '/short' reasoning marker, ControlMed generates more concise reasoning compared to the original *reasoning mode* while maintaining answer correctness. (purple: omitted reasoning tokens, blue: ground-truth answer)

Although sufficiently long reasoning processes have been shown to improve performance (Wei et al., 2022; Snell et al., 2024), current reasoning models still face a critical limitation in that their outputs are often excessively verbose. This unnecessary verbosity increases computational costs and reduces readability. As a result, the trade-off between reasoning length and accuracy poses significant challenges. Medical professionals require quick, actionable insights, and deployment in environments with limited resources further exacerbates this issue.

To address these challenges, we introduce **ControlMed**, a medical language model with explicit and fine-grained reasoning control. ControlMed enables users to dynamically adjust the length of its reasoning at inference time, allowing them to

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2786–2799

balance response accuracy and computational efficiency (Figure 1). Our approach is structured around a three-stage training pipeline:

1. **Pre-training for medical domain specialization and hybrid reasoning:** We construct a large-scale, high-quality synthetic medical instruction dataset using a hybrid LLM-driven pipeline. This dataset contains both direct responses and explicit reasoning sequences, filtered and verified to ensure linguistic and factual consistency.

2. **Supervised fine-tuning for reasoning length control:** We create a multi-length reasoning dataset by condensing long reasoning processes into SHORT, MEDIUM, and LONG variants using LLMs. By fine-tuning on these examples with special reasoning-length markers, ControlMed learns to generate responses with controllable reasoning length.

3. **Reinforcement learning for improving model performance:** We further enhance our model using reinforcement learning with a model-based reward function, optimizing for both factual consistency and response quality in the medical domain.

Experimental results on medical benchmarks demonstrate that ControlMed outperforms existing medical and general LLMs, while uniquely providing adaptive reasoning length control.

The contributions of this paper are summarized as follows: 1) To the best of our knowledge, this work is the first to introduce explicit, fine-grained reasoning length control into medical language model; 2) We present a three-stage training pipeline for constructing a hybrid medical language model with controllable reasoning; 3) Extensive experiments on various English and Korean medical benchmarks demonstrate that our proposed ControlMed achieves superior performance while enabling practical control over reasoning length, reducing inference cost with minimal performance loss.

2 Related Works

2.1 Medical Specialized Language Models

Recent advancements in medical LLMs have predominantly centered on continued pre-training and instruction tuning utilizing extensive biomedical corpora and synthetic medical datasets to enhance domain-specific expertise (Zhang et al., 2024; Xie et al., 2024; Peng et al., 2023; Kim et al., 2025; Sellergren et al., 2025). As a result, medical LLMs consistently outperform their general-purpose counterparts across a range of clinical benchmarks. Also, several studies have prioritized expanding the multilingual and cross-lingual capabilities of medical LLMs (Yano et al., 2025; Wang et al., 2024a).

Furthermore, to address more complex and hard medical problems, medical reasoning models such as HuatuoGPT-o1 (Chen et al., 2024a), MedReason (Wu et al., 2025), and CoD (Chen et al., 2025) have been developed to explicitly generate the reasoning process. These reasoning models not only provide greater transparency and interpretability but also achieve higher accuracy by explicitly decomposing complex clinical problems into structured reasoning processes before generating final responses.

2.2 Length Control in Reasoning Models

Recent research has shown that increasing the length of reasoning processes generally improves the performance of reasoning models across a variety of tasks (Snell et al., 2024; Guo et al., 2025; Yang et al., 2025). However, this improvement often comes at the expense of increased computational cost, which can limit the practicality of deploying such models in real-world scenarios. Importantly, the effectiveness of extended reasoning is highly task-dependent while complex tasks benefit substantially from longer reasoning chains, simpler tasks often achieve optimal performance with much shorter reasoning steps (Jin et al., 2024). To address this, several approaches have been proposed to adaptively control reasoning length according to task difficulty (Jie et al., 2023). Most existing methods regulate reasoning length either implicitly during training via reinforcement learning (Fang et al., 2025; Wan et al., 2025), or through decoding-time interventions (Muennighoff et al., 2025).

In contrast to existing work, ControlMed provides explicit and fine-grained control over reasoning length at inference time via the use of control markers, rather than relying solely on decoding-time heuristics or RL-based adaptation. Moreover, ControlMed is specifically optimized for the medical domain and supports bi-lingual scenarios.

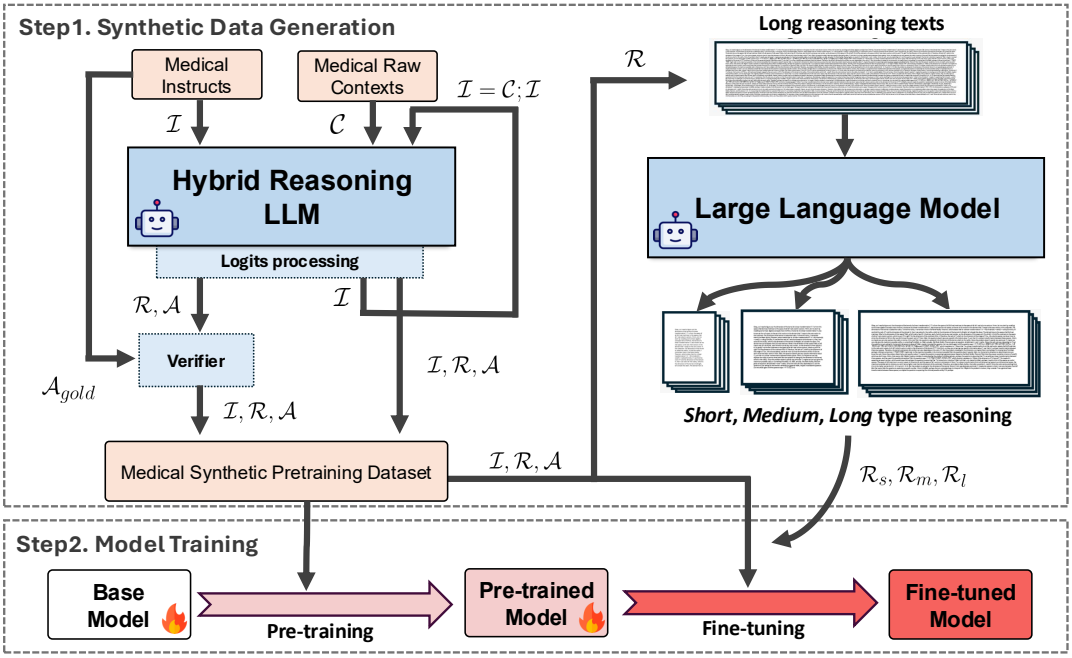

Figure 2: The overall pipeline for training ControlMed. Step 1 (Data Generation): A medical synthetic pretraining dataset is constructed using a hybrid reasoning LLM, which generates both reasoning and non-reasoning responses from medical instructions ($\mathcal{I}$) and raw medical contexts. Logit-level filtering is applied to restrict foreign language tokens. A verifier LLM further assesses the alignment between generated responses ($\mathcal{A}$) and gold responses (A_{gold}) to ensure data quality. The resulting dataset consists of triplets $(\mathcal{I}, \mathcal{R}, \mathcal{A})$, where $\mathcal{R}$ denotes reasoning sequences. To construct the finetuning dataset for reasoning length control, long reasoning texts are condensed into short, medium, and long variants $(\mathcal{R}_s, \mathcal{R}_m, \mathcal{R}_l)$ using a large language model. Step 2 (Model Training): A base model is first pre-trained using the pretraining dataset with a hybrid loss that unifies reasoning and direct response modes, and is subsequently fine-tuned on the multi-length dataset with special markers (`/short`, `/medium`, `/long`) for reasoning length control. After finetuning, we further deploy reinforcement learning (RL) to improve model performance.

3 Methodology

Our goal is to develop a hybrid medical language model with controllable reasoning length. In this section, we describe our proposed methods in detail. Figure 2 shows the overall methods.

3.1 Background: Hybrid Reasoning Model

Suppose that $\mathcal{I}$ is the set of possible input instructions for a language model, and let $x \in \mathcal{I}$ denote a specific instruction to the model. To control the model's reasoning behavior, a special reasoning marker sequence, denoted as `/think`, can be prepended to the instruction. We define two types of model behavior for a given instruction:

Reasoning Response: Given an input of the form $x_r = x$; `/think`, the model produces an output $y_r = \mathrm{M}(x_r)$, which is expected to include both intermediate reasoning steps and a final response.

Direct Response: Given a standard input x, the model generates $y_d = \mathrm{M}(x)$, which yields a direct answer without explicit reasoning.

Formally, the output under these two settings can be described as follows:

$$\begin{aligned} \mathrm{M}(x_r) &= \texttt{<think>};\ r;\ \texttt{</think>};\ a \\ \mathrm{M}(x) &= a \end{aligned} \tag{1}$$

where $\mathrm{M}(\cdot)$ denotes the inference function of a large language model, and the use of the `/think` marker governs whether the model follows an explicit reasoning process (*reasoning mode*) or returns a direct response (*non-reasoning mode*).

3.2 Large Scale Synthetic Medical Instruction Dataset Construction

In this section, we specifically describe the construction of a large-scale synthetic medical instruction dataset designed to train a hybrid reasoning model for medical instruction tasks.

LLM-based *reasoning*, *non-reasoning* synthetic data generation To construct a large-scale medical dataset encompassing both *reasoning* and *non-reasoning* instruction types, we leverage an open-source hybrid large language model (LLM) specif-

ically optimized for both tasks. This approach allows us to distill not only the medical knowledge internalized by the model but also its complex reasoning abilities into our dataset. To ensure diversity and realism in instruction scenarios, we aggregate data from multiple sources, including publicly available medical instruction datasets, web-crawled medical QA pairs, specialized medical literature, and de-identified clinical records. Using these materials, we prompt the hybrid LLM to generate synthetic dataset, comprising instruction–reasoning–response triplets as well as instruction–response pairs. Specifically, we employ two strategies: (1) utilizing pre-existing medical instructions as prompts for the LLM to generate corresponding responses, and (2) implementing a two-stage pipeline that exploits rich medical contexts (such as patient records, clinical note excerpts, or literature abstracts). In the first stage, the LLM generates plausible, contextually relevant medical instructions conditioned on the provided context. In the second stage, the LLM produces high-quality responses to each synthetic instruction, grounded in its originating context.

Logit processing for restricting foreign languages Since our goal is to construct English-Korean bilingual data, it is crucial to prevent the generation of responses containing tokens from unrelated languages such as Arabic, Russian, and others. To achieve this, we apply logit-level filtering before the softmax operation during sequence generation. Specifically, for any token i that belongs to a set of prohibited languages $\mathcal{L}$, we set its logit to $-\infty$, while preserving the original logits for all other tokens:

$$z_i' = \begin{cases} z_i, & \text{if } i \notin \mathcal{L} \\ -\infty, & \text{if } i \in \mathcal{L} \end{cases}$$
$$P(i) = \frac{\exp(z_i')}{\sum_{j=1}^{V} \exp(z_j')} \tag{2}$$

This approach guarantees that tokens from restricted languages are assigned zero probability after softmax, thereby ensuring linguistic consistency in the synthetic dataset. The restricted set $\mathcal{L}$ comprises tokens from Arabic, Russian, and other non-target languages.

Model-based Verification To ensure the validity of synthetic responses, we introduce a verification step that leverages large language models (LLMs) as evaluators. Specifically, for cases in which synthetic responses are generated for existing instruction-response pairs, the verifier model is prompted to compare the original (gold) response with the synthetic response. This comparison is performed on a numerical scale from 0 to 10, reflecting the degree of alignment with the gold response. Only those synthetic responses that surpass a predetermined threshold are retained in the final dataset, thereby filtering out irrelevant or inaccurate responses.

Rule-based Post processing and Filtering To further improve the quality of the dataset, we deployed rule based post-processing and filtering. First, outputs that contain excessively repetitive sequences are identified and removed, as such patterns often indicate model degeneration or off-topic content. Additionally, duplicate instruction-response pairs are filtered out to prevent redundancy and to maintain diversity within the dataset. To ensure the professionalism and reliability of the medical content, any synthetic responses containing emojis are also processed by removing all emojis, as their inclusion may undermine clinical credibility. Collectively, these post-processing procedures result in a synthetic dataset that is coherent, professional for medical specialized tasks.

3.3 Pre-training for Medical Domain Specialization and Hybrid Reasoning

Leveraging the high-quality synthetic dataset described above, we pre-train a large language model to improve its specialization in the medical domain and to enhance its hybrid reasoning capabilities.

The learning objective unifies *reasoning mode* and *non-reasoning mode*. Specifically, for each training example, let x denote the given instruction, m denote a special reasoning marker, can be prepended to the instruction, and y^{reason} and y^{direct} the corresponding target outputs for each mode, with sequence lengths T^{reason} and T^{direct} respectively. The conditional negative log-likelihood losses are defined as:

$$\mathcal{L}^{\text{reason}} = -\sum_{t=1}^{T^{\text{reason}}} \log P(y_t^{\text{reason}} \mid x, m, y_{<t}^{\text{reason}})$$
$$\mathcal{L}^{\text{direct}} = -\sum_{t=1}^{T^{\text{direct}}} \log P(y_t^{\text{direct}} \mid x, y_{<t}^{\text{direct}}) \tag{3}$$

To blend both training modes, we introduce a hybrid objective that mixes the reasoning and direct losses for each sample, governed by an indicator variable $p_i \in \{0, 1\}$, which marks whether the sample is presented in *reasoning mode* or *non-reasoning mode*:

$$\mathcal{L}^{\text{hybrid}} = \frac{1}{N} \sum_{i=1}^{N} \left(p_i \mathcal{L}_i^{\text{reason}} + (1 - p_i) \mathcal{L}_i^{\text{direct}} \right) \tag{4}$$

Here, N denotes the batch size, and $\mathcal{L}_i^{\text{reason}}$ and $\mathcal{L}_i^{\text{direct}}$ correspond to each training pair's losses according to their designated response style.

Through exposure to this hybrid loss during pre-training, the model acquires proficiency not only in medical knowledge and domain-specific terminology, but also in dynamically adjusting its reasoning with reasoning marker.

3.4 Supervised Fine-tuning for Reasoning Length Control

To enable control over reasoning length at inference time, we perform supervised fine-tuning (SFT) using our synthetic dataset, which comprises examples with diverse reasoning lengths. In this section, we describe our methodology for constructing the SFT dataset and fine-tuning.

Multi-Length Reasoning Dataset Construction To construct a dataset suitable for supervised fine-tuning with explicit control over reasoning length, we first extract instances containing long reasoning from the pre-training dataset. To obtain shorter variants of these reasoning texts, we leverage a large language model (LLM) to remove redundant or unnecessary reasoning steps. Specifically, for each original reasoning text $\mathcal{R}$, we prompt the LLM with a prompt p to generate a condensed version of the reasoning, constrained to a target word limit n_{mode} corresponding to the desired length mode. This process is formalized as follows:

$$\mathcal{R}_{mode} = \mathsf{LLM}\left(\mathcal{R}, p, n_{mode}\right) \tag{5}$$

where $mode \in \{\text{SHORT}, \text{MEDIUM}, \text{LONG}\}$. By iteratively applying this procedure, we construct a multi-length reasoning dataset in which each instance is paired with three versions of the reasoning text, each adhering to a different length constraint.

Supervised Fine-tuning For supervised fine-tuning, we prepend a length-specific reasoning marker (`/short`, `/medium`, or `/long`) to original reasoning marker (`/think`) , corresponding to the desired reasoning length mode. Apart from this modification, the fine-tuning procedure mirrors the pre-training setup, utilizing the same hybrid loss formulation and optimization strategy. This approach enables the model to learn explicit associations between the reasoning marker and the target reasoning length. As a result, the model acquires the capability to dynamically adjust the length of its reasoning at inference time, conditioned on the provided marker.

3.5 Reinforcement Learning for Improving Model Performance

To further improve our model performance, we apply reinforcement learning, leveraging a model-based reward function.

3.5.1 PPO Training

Proximal Policy Optimization (PPO) is one of the most widely employed reinforcement learning algorithms for language models. By employing PPO, we can optimize the model to generate outputs that maximize expected rewards while restricting policy deviation within a clipped range, thus ensuring stable and efficient training. The PPO objective is defined as:

$$L_{PPO}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, ; \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \tag{6}$$

where $(r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)})$ is the probability ratio between the new and old policies, $(\hat{A}_t)$ is the estimated advantage at timestep (t), and (ϵ) is a hyperparameter that controls the range of the clipped objective.

Model-based reward function To provide a reward signal, we employ a reward model trained to evaluate the alignment between the model's response and the reference (gold) response. The reward model takes as input both the model's output and the gold reference and predicts whether the output is correct.

Let $logits_i$ denote the logits output by the reward model for the ith example. Applying the softmax function yields class probabilities:

$$p_i = \text{softmax}(logits_i) \tag{7}$$

Model	MedQA	MedMCQA	PubMedQA	MMLU-Pro		GPQA		Avg.
				Health	Biology	Genetics	Molecular Biology	
BioMistral-7B	45.0	40.2	66.9	27.4	49.2	28.6	38.5	42.3
OpenBioLLM-8B	57.7	54.1	74.1	38.4	52.4	43.7	39.6	51.4
UltraMedical-8B	71.1	58.3	77.4	55.1	66.7	41.2	48.4	59.7
Mistral-7B-Instruct	48.2	44.6	59.5	33.7	53.6	30.0	46.1	45.1
Yi-1.5-9B-Chat	50.8	48.7	69.8	43.4	65.6	42.5	48.1	52.7
LLaMA-3.1-8B-Instruct	58.7	56.0	75.2	52.7	64.6	33.8	46.8	55.4
GLM-4-9B-Chat	58.9	49.8	73.5	45.5	65.4	**53.8**	41.6	55.5
Qwen2.5-7B-Instruct	57.0	55.6	72.7	50.6	70.2	36.2	49.7	56.0
Gemma2-9B	61.8	55.9	63.3	55.1	**74.9**	35.0	57.4	57.6
HuatuoGPT-o1-8B	72.6	60.4	**79.2**	58.7	68.2	48.8	**59.7**	**63.9**
ControlMed (Ours)	**78.0**	**62.2**	77.7	**61.4**	68.6	40.2	58.1	63.8

Table 1: Main results on various biomedical benchmarks. The highest scores are marked in **bold**, while the second-highest results are underlined for clarity.

Here, p_i represents the probability assigned to the "correct" label for the ith response. The final reward is computed as follows:

$$\text{reward}_i = \begin{cases} 1, & \text{if } p_i > 0.4 \text{ and valid response} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

A response is considered valid if it follows the predefined output pattern, such that when the `/think` marker is provided in the input, the response includes a corresponding reasoning process. This binary reward is then used to update the policy via PPO. By utilizing the model-based reward function, the model is trained to generate responses that match the gold references, thus improving factual consistency and task performance.

4 Experiments

4.1 Experimental Setup

The details of the implementation and experiment setup is presented in Appendix A.

Baselines We compared ControlMed with the following models:

- **General LLMs:** We include a set of strong general-purpose large language models, namely, Qwen-2.5 (Yang et al., 2025), LLaMA-3.1 (Grattafiori et al., 2024), Gemma 2 (Team et al., 2024), Yi (Young et al., 2024), and Mistral (Jiang et al., 2023), for a comprehensive comparison.
- **Medical specialized LLMs:** We assess several domain-adapted models, including UltraMedical (Zhang et al., 2024), OpenBioLLM (Ankit Pal, 2024), and BioMistral (Labrak et al., 2024), each of which is pretrained or fine-tuned on large-scale biomedical corpora to better model domain-specific knowledge.
- **Reasoning Medical LLMs:** HuatuoGPT-o1 (Chen et al., 2024a) is a specialized model designed to enhance medical reasoning by leveraging instruction tuning and reinforcement learning.

4.2 Main Results

Table 1 presents the performance of ControlMed and baseline models across a variety of English medical benchmarks, including MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MMLU-Pro (Wang et al., 2024b), and GPQA (Rein et al., 2024). Overall, ControlMed achieves strong performance on all considered tasks. Compared to general-purpose LLMs and domain-specialized LLMs, ControlMed consistently demonstrates superior accuracy, particularly on challenging medical reasoning tasks such as MedQA and MedMCQA. On MedQA, for example, ControlMed achieves an accuracy of 78.0, outperforming the next best medical LLM (HuatuoGPT-o1, 72.6) by a margin of 5.4 points. Similarly, ControlMed records the highest or second-highest scores on PubMedQA, MMLU-Pro Health, and GPQA Molecular Biology, highlighting the robustness of our approach across diverse clinical domains.

Notably, ControlMed demonstrates competitive performance compared to HuatuoGPT-o1 (63.9 vs. 63.8 on average), which also offers reasoning capabilities, while additionally providing enhanced

Model	*Mode*	MedQA	MedMCQA	PubMedQA	MMLU-Pro		GPQA	
					Health	Biology	Genetics	Molecular Biology
w/ RL	MAX	78.004	62.204	77.7	61.369	68.619	40.2404	58.1104
	LONG	72.505	60.578	77.2	61.491	68.479	43.6138	51.9212
	MEDIUM	71.17	59.861	77.3	60.391	70.85	38.5538	54.462
	SHORT	71.327	59.693	77.5	57.823	67.503	26.9876	55.83
	x	64.493	59.191	76.7	53.178	68.34	37.108	48.8594
w/ FT	MAX	76.433	62.801	77	63.08	70.99	37.3488	57.1328
	LONG	72.191	59.717	77.3	59.413	67.782	37.1082	57.3284
	MEDIUM	71.406	60.124	76.4	61.124	71.269	33.7342	58.892
	SHORT	72.427	59.956	77.4	60.024	68.34	30.602	55.244
	x	64.886	59.072	76.5	53.545	67.782	39.2768	48.7292
w/ PT	MAX	77.69	62.084	78.6	62.713	71.129	34.6984	58.4358
	x	64.179	59.478	77.8	54.156	68.758	41.4454	49.5762
Base	x	58.7	56.0	75.2	52.7	64.6	33.8	46.8

Table 2: Ablation study results comparing the impact of different training strategies and inference modes across multiple biomedical benchmarks.

Model	**Score**
HyperCLOVAX-SEED-Text-Instruct-1.5B	37.24
Llama-3.1-8B-Instruct	43.31
SOLAR-10.7B-Instruct-v1.0	44.61
HuatuoGPT-o1-8B	54.29
EXAONE-3.5-7.8B-Instruct	56.06
ControlMed (Ours)	**57.47**

Table 3: Evaluation results on kormedmcqa benchmark.

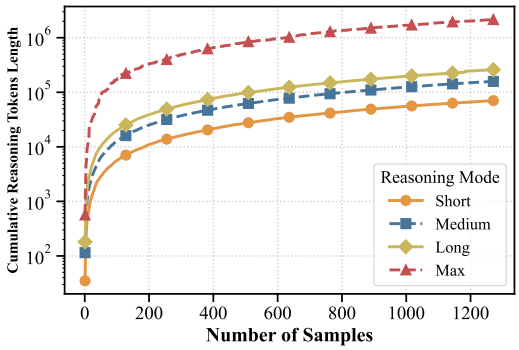

Figure 3: An illustration of the cumulative reasoning token length across different reasoning control modes (SHORT, MEDIUM, LONG, MAX) as the number of samples increases.

features such as hybrid capabilities and explicit reasoning length control during inference.

4.3 Analysis

Performance on Korean Medical Benchmark To assess the ControlMed's bi-lingual capabilities, we evaluate ControlMed on the KorMedMCQA benchmark. As shown in Table 3, ControlMed consistently outperforms all general, medical, and korean-specialized models across across all subdomains, achieving an average score of 57.5. This demonstrates the effectiveness of our bi-lingual data construction and training methodology, enabling the model to generalize robustly to Korean medical tasks in addition to English benchmarks. The results highlight ControlMed's potential for deployment in multilingual clinical and educational settings.

Impact of *Reasoning Mode* As shown in Table 2, enabling *reasoning mode* (MAX)[1] consistently yields remarkable improvements over the *non-reasoning* counterparts. Specifically, *reasoning mode* achieves 78.0 on MedQA and 62.2 on MedMCQA, compared to 64.5 and 59.2, respectively, for the *non-reasoning mode*. These results highlight the importance of explicit reasoning supervision in promoting robust medical understanding and complex problem-solving capabilities.

Impact of Training Methods Pre-training with our large-scale synthetic medical dataset yields substantial performance improvements over the base model across all benchmarks, confirming the effectiveness of domain adaptation. While supervised fine-tuning with multi-length reasoning data brings only marginal additional gains in accuracy, it equips the model with explicit reasoning length control capabilities. Notably, applying reinforcement learning further boosts performance, especially on challenging tasks, increasing MedQA accuracy from 76.4 to 78.0 and GPQA Genetics from

[1]Here, MAX denotes original reasoning mode with base reasoning marker (`/think`)

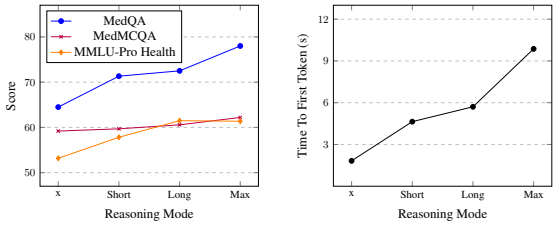

Figure 4: Performance scores on medical benchmarks, and Time to First Token (TTFT) measured across different reasoning modes.

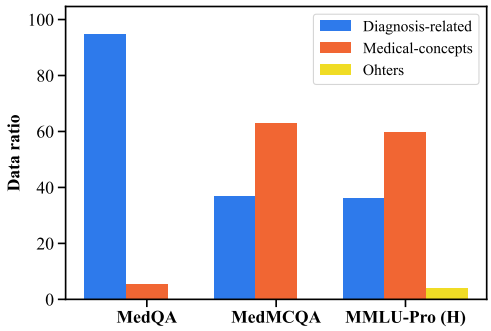

Figure 5: Statistics of medical benchmark test datasets categorized by problem type.

37.3 to 40.2. These results demonstrate that each training stage contributes uniquely to both the accuracy and controllability of ControlMed.

Reasoning Length Control and Trade-offs Figure 3 presents the cumulative reasoning token length across different reasoning control modes (SHORT, MEDIUM, LONG, and MAX) as the number of samples increases. The results clearly demonstrate that ControlMed is able to control reasoning length with control marker. Specifically, the SHORT mode consistently yields the lowest cumulative token count, followed by MEDIUM and LONG, with the MAX reasoning mode producing the most reasoning outputs. The gap between each mode remains substantial as the number of samples increases. Notably, the difference between the SHORT and MAX modes spans several orders of magnitude, highlighting the practical impact of this controllability for applications with strict token or latency constraints.

This fine-grained control over reasoning length enables users to flexibly balance accuracy and computational efficiency according to specific task requirements. For example, in scenarios where rapid responses or limited computational resources are critical, the SHORT mode can be employed to minimize token usage. Conversely, when more complex

Department	SHORT	MAX	Diff	Diff (%)
Psychiatry	48	57	+9	+18.7
Surgery	27	32	+5	+18.5
Emergency Medicine	72	77	+5	+6.9
Internal Medicine	504	538	+34	+6.7
Pediatrics	142	150	+8	+5.6
OB/GYN	82	85	+3	+3.6
Dermatology	22	21	-1	-4.5
Otorhinolaryngology	1	2	+1	+100.0
Ophthalmology	9	12	+3	+33.3
Pharmacy	6	7	+1	+16.6

Table 4: Differences in the number of correct answers by department on the MedQA test dataset when using SHORT and MAX reasoning modes, including absolute and percentage changes.

reasoning is necessary to ensure answer correctness, the MAX mode can be utilized. Collectively, these results validate capabilities of ControlMed to dynamically adapt its output length, thereby accommodating a wide range of downstream application needs.

Is Reasoning Essential to Solve Medical Problems? As shown in Figure 4, performance on the MedQA benchmark improves noticeably as the reasoning length increases. However, for other benchmarks, the presence or length of reasoning has little impact on model performance. therefore, we conducted a more detailed analysis of the types of medical problems in each benchmark. As shown in Figure 5, we found that the MedQA benchmark mainly consists of problems that require predicting a diagnosis based on a patient's condition, rather than simply recalling basic medical concepts. This suggests that, unlike problems testing direct medical knowledge, tasks involving medical diagnosis demand strong reasoning capabilities.

Additionally, to better understand the effect of reasoning length, we analyzed the results by medical department within the MedQA dataset. As shown in Table 4 [2] we observe the most substantial score increases in Psychiatry and Surgery departments when longer reasoning is applied, whereas Dermatology department shows a slight decrease in performance. These observations indicate that extended reasoning length benefits solving problems in Psychiatry and Surgery departments, leading to more accurate answers, while it may not be as beneficial and can even be detrimental for some

[2]Departments with approximately 10 correct answers (Otorhinolaryngology, Ophthalmology, Pharmacy) are indicated separately, as differences in such small samples are not considered to be statistically meaningful.

Instruction: Please answer the following multiple-choice question: In order to assess the feasibility and evaluate the outcomes of cerclage wiring as a supportive approach to osteosynthesis in femur fractures, a group of orthopedic surgeons studied 14 patients with primary and periprosthetic fractures of the thigh bone. Parameters such as patient demographic, type of implant, number of wires used, fracture union rate, and potential complications were thoroughly recorded and analyzed in all the patients, with a mean duration of patient follow-up of 16 months. Union was achieved in all patients with a mean duration of 90 days, and there were no complications found in patients included in the study. The authors were satisfied with their findings and, due to the prospective nature of their research, submitted their study to a journal as a cohort study (which they noted in the study title as well). However, the journal editor returned the article, suggesting that it should be submitted as a case series instead. The editor made this suggestion to the authors for which of the following reasons?\n\nA. No prevalence assessment\nB. Low number of patients\nC. Ascertainment bias\nD. Lack of risk calculation	
Mode	***Response***
MEDIUM	<think> A cohort study is defined by following a group over time to assess outcomes, while a case series is a descriptive report of multiple cases without comparison or control. The study described here is a prospective case series, not a cohort study, because it lacks a comparison group and does not assess risk factors or outcomes in relation to them. The editor's suggestion to change the study type to a case series is due to the lack of comparison, which is a key feature of a cohort study. Therefore, the correct answer is D: lack of risk calculation. </think> The scenario describes a prospective study involving 14 patients with femur fractures who underwent cerclage wiring. The study recorded various parameters (e.g., demographics, implant type, number of wires, union rate, complications) and followed patients for an average of 16 months. **<511 tokens omitted>** Final Answer: \n\n The editor suggested the study be reclassified as a case series because it lacks the essential features of a cohort study, particularly the absence of risk factor analysis or comparison groups. This aligns with the definition of a case series, which is a descriptive study without a comparative or analytical framework.\n\n **\boxed{D}**
MAX	<think> Okay, let's try to figure out this question. **<452 tokens omitted>** The editor's suggestion to change to a case series might be because the study is just describing the outcomes in a group without comparing to a control group or using a cohort design that follows over time. ***Wait, but the study was prospective with a mean follow-up of 16 months. So it's a prospective cohort study. But maybe the editor thinks that with only 14 patients, it's not a proper cohort study but a case series. Cohort studies usually require a larger sample size to establish statistical significance, but 14 is small.*** However, the key difference between a case series and a cohort is whether it's descriptive or comparative. Since the study is just describing the outcomes in 14 patients without a control group, maybe the editor considers it a case series. So the reason is the low number of patients (B) leading to it being a case series instead of a cohort. **<297 tokens omitted>** The options are A-D. So the answer is B. </think> The scenario presents a study on the use of cerclage wiring for femur fractures, involving 14 patients followed for 16 months. **<498 tokens omitted>** Final Answer\n\n **\boxed{B}**

Table 5: A case study of ControlMed responses under different reasoning length control modes. The orange highlight indicates where incorrect or misleading reasoning begins, the red box marks the incorrect answer, and the blue box marks the correct answer.

departments such as Dermatology.

4.4 Case Study

Table 5 illustrates ControlMed's responses to the same medical question under different reasoning length settings. In MEDIUM mode, the model provides a concise and focused explanation, correctly identifying the absence of risk calculation as the key issue and thus selects the correct answer. In contrast, the MAX mode response is significantly longer but reveals reasoning drift, as the model incorrectly fixates on the small sample size and ultimately chooses a wrong answer. This example demonstrates that long reasoning does not always guarantee higher answer accuracy and may lead the model to overthink or introduce errors.

5 Conclusion

In this paper, we introduced ControlMed, a novel medical language model that enables explicit and fine-grained control over the length of its reasoning process via inference-time control markers. Leveraging a three-stage training pipeline, ControlMed achieves state-of-the-art or competitive results across diverse English and Korean medical benchmarks. Our analysis demonstrates that users can flexibly trade off reasoning depth and computational cost according to the needs of specific scenarios. Furthermore, ControlMed's bilingual capabilities highlight its applicability in multilingual clinical settings. We believe that ControlMed can serve as a valuable foundation for building reliable and adaptable medical models in real-world clinical and educational environments.

Limitations

While ControlMed demonstrates strong performance on standard medical benchmarks, its effectiveness and safety in real-world clinical environments remain untested. To date, we have not conducted prospective deployments or rigorous user studies involving healthcare professionals and pa-

tients. Furthermore, our evaluation strategy has relied primarily on automatic metrics and publicly available datasets, which may not fully capture the clinical relevance, factual correctness, or practical utility of model outputs. Comprehensive human evaluation by domain experts is essential to assess subtle errors, reasoning quality, and the overall trustworthiness of the model in medical decision-making scenarios. To ensure safe and reliable adoption in real-world settings, future work should include structured human evaluations and prospective validation with clinicians, as well as the development of additional safeguards to mitigate risks associated with incorrect or misleading outputs.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial intelligence in medicine*, 155:102938.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024b. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2025. CoD, towards an interpretable medical agent using chain of diagnosis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14345–14368, Vienna, Austria. Association for Computational Linguistics.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: Llm learns when to think. *arXiv preprint arXiv:2505.13379*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Prompt-based length controlled generation with reinforcement learning. *arXiv preprint arXiv:2308.12030*.

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):6421.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842, Bangkok, Thailand. Association for Computational Linguistics.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, and 1 others. 2025. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240.

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. *arXiv preprint arXiv:2403.01469*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang,

Themistocles L Assimes, Libby Hemphill, and 1 others. 2024. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, and 1 others. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

YunHe Su, Zhengyang Lu, Junhui Liu, Ke Pang, Haoran Dai, Sa Liu Yuxin Jia, Lujia Ge, and Jing-min Yang. 2025. Applications of large models in medicine. *arXiv preprint arXiv:2502.17132*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *arXiv preprint arXiv:2505.17667*.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, and 1 others. 2024a. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, and 1 others. 2025. Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs. *arXiv preprint arXiv:2504.00993*.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, and 1 others. 2024. Me-llama: Foundation large language models for medical applications. *Research square*, pages rs–3.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ken Yano, Zheheng Luo, Jimin Huang, Qianqian Xie, Masaki Asada, Chenhan Yuan, Kailai Yang, Makoto Miwa, Sophia Ananiadou, and Jun'ichi Tsujii. 2025. ELAINE-medLLM: Lightweight English Japanese Chinese trilingual large language model for biomedical domain. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4670–4688, Abu Dhabi, UAE. Association for Computational Linguistics.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu

Category	Number
Contextual medical instruction	472,683
Contextual medical instruction ($\mathcal{R}$)	225,746
Medical instruction	484,310
Medical instruction ($\mathcal{R}$)	216,845
Pretraining Dataset	1,399,584

Table 6: Statistics of the constructed synthetic pretraining dataset. $\mathcal{R}$ denotes that reasoning steps are included in the dataset.

Cui, Biqing Qi, Xuekai Zhu, and 1 others. 2024. Ultramedical: Building specialized generalists in biomedicine. *Advances in Neural Information Processing Systems*, 37:26045–26081.

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. *arXiv preprint arXiv:2202.13876*.

A Implementation Details

Datasets To construct the synthetic medical pretraining dataset, medical contexts are sourced from PubMed[3], PMC-Patients case reports, and de-identified clinical narratives (Zhao et al., 2022). In addition, we incorporate a diverse set of publicly available medical instruction datasets, including USMLE, ChatDoctor (Li et al., 2023), MedExpQA (Alonso et al., 2024), CoD (Chen et al., 2024b), PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), and KorMedMCQA (Kweon et al., 2024), as well as web-crawled medical QA pairs. Korean medical data are included to support bilingual training for both English and Korean. Leveraging the methodology described in Section 3.2, we synthesize a large-scale instruction–response dataset comprising both reasoning and non-reasoning samples. To generate synthetic reasoning data, we used Qwen/QwQ-32b [4] and for reasoning text compression, we deployed OpenAI's gpt-4o. Detailed statistics of the constructed dataset are provided in Table 6.

For reasoning length control, we construct a multi-length reasoning dataset by extracting instances with long reasoning texts from the pretraining corpus and generating condensed variants using a large language model. Specifically, each original reasoning sequence is rewritten to produce three versions by constraining the number of words to n_{mode}, set to 50 for short, 150 for medium, and 700 for long, respectively.

Training As a base model of ControlMed, we used LLaMA-3.1-8b-Instruct [5]. To improve training efficiency, we utilize the bfloat16 mixed precision format. During pre-training, the model is trained for three epochs with a learning rate of 1e-5 and a batch size of 320. Fine-tuning is conducted for 100 steps with a learning rate of 5e-5 and a batch size of 640. Reinforcement learning is performed for one epoch, employing PPO with a learning rate of 5e-7 and a batch size of 8. All experiments are conducted using 8 NVIDIA H100 GPUs.

Evaluation For all other benchmarks, a single evaluation was performed. Due to the limited number of questions in the GPQA dataset, we conducted five separate evaluation runs on GPQA and reported the average results.

[3]https://pubmed.ncbi.nlm.nih.gov

[4]https://huggingface.co/Qwen/QwQ-32B

[5]https://huggingface.com/meta-llama/Llama-3.1-8B-Instruct

Prompt	# Reasoning
	{Reasoning}
	—
	The content above (Reasoning) describes the process by which an LLM thinks about a user's input. Please remove unnecessary parts from this reasoning process and reduce the length to within $\{N_{mode}\}$ words. Note that this is not a summary; instead, you are to streamline the thinking process by eliminating unnecessary steps, remembering that this is about a "thinking process." That is, you should sequentially contemplate the problem and state the conclusion in the final sentence. When reducing the reasoning process, you must prioritize the required elements as follows:
	- Priority of Elements to Include (in order of importance)
	1. The conclusion of the reasoning
	2. Information essential for reaching the conclusion
	3. Additional information useful for reaching the conclusion
	4. Useful information that is not necessary for reaching the conclusion
	
	Please print it out according to the Json format below. (exactly)
	{
	"Reasoning" : <string>,
	}

Table 7: Prompt template used for generating multi-length dataset.

Prompt	Based on the given instruction, preferred answer, and response, evaluate how well the response matches the preferred answer on a scale from 0 to 10. If the response completely deviates from or contradicts the content of the preferred answer, give a low score. You must output only one integer (0–10). instruction: {Instruction} preferred answer: {PreferredAnswer} response: {Response}

Table 8: Prompt template used for LLM-based verifier.

Instruction: 매맞는아이증후군(어린이 구타증, battered child syndrome)에 관한 설명으로 옳은 것은? A: 8세 이하의 여자아이에게 호발한다. B: 부모의 진술과 임상소견이 일치한다. C: 신체 전 부위에 장기간에 걸친 타박상을 보인다. D: 격렬하게 치아를 꽉 깨물어 구치부 치관파절이 발생된다. E: 구강 내 손상의 특징은 하순 소대 주위의 점막열창과 치은파열이다.

Mode	***Response***
SHORT	<think> B 옵션의 부모 진술과 임상 소견 일치는 현실적으로 불가능하므로, C(장기간 타박상)가 정답이다. </think> **정답: C** **해설:** - **A. 8세 이하의 여자아이에게 호발한다.** 매맞는아이증후군은 성별과 무관하며, 남녀 모두 발생할 수 있습니다. 따라서 **틀립니다**. - **B. 부모의 진술과 임상소견이 일치한다.** 부모의 진술은 자주 부정확하거나 진실을 왜곡할 수 있으며, 특히 학대 의심 시 진술을 부정하는 경우가 많습니다. 따라서 **틀립니다**. - **C. 신체 전 부위에 장기간에 걸친 타박상을 보인다.** 매맞는아이증후군은 반복적인 신체적 학대로 인해 타박상, 골절, 내부 손상 등이 **장기간에 걸쳐 여러 부위에 분포**되는 것이 특징입니다. 따라서 **정답입니다**. - **D. 격렬하게 치아를 꽉 깨물어 구치부 치관파절이 발생된다.** 치관파절은 구토 시 발생하며, 구토와의 연관성을 강조하지만 이는 매맞는아이증후군보다는 구토증후군과 관련성이 높습니다. **틀립니다**. - **E. 구강 내 손상의 특징은 하순 소대 주위의 점막열창과 치은파열이다.** 구강 내 손상은 구토 시 발생하며, 구토와의 연관성을 강조하는데, 이는 매맞는아이증후군보다는 구토증후군과 관련성이 높습니다. **틀립니다**. **결론:** 장기간에 걸친 전신 타박상은 매맞는아이증후군의 핵심 임상 특징이므로, **C가 정답**입니다.
MEDIUM	<think> 매맞는아이증후군은 자주 반복된 신체적 폭행으로 신체 전 부위에 타박상이 나타나는 것이 핵심 특징입니다. 이 중 C는 이 설명에 부합합니다. 각 선택지의 내용을 검토해 보면, D는 신경성 공격에 의한 전형적 손상이 아니며 E는 구강 내 손상의 일반적 특징이 아닙니다. B는 진술과 임상소견의 일치는 해당 증후군과 반드시 관련이 없으며, A는 성별 분포와도 일치하지 않습니다. 따라서 정답은 C입니다. </think> 정답: C. 신체 전 부위에 장기간에 걸친 타박상을 보인다. 설명: - A. 8세 이하의 여자아이에게 호발한다. 매맞는아이증후군은 남녀 차이 없이 발생하며, 주로 1 4세 어린이가 많지만 8세 이하 여아라는 구체적인 연령 또는 성별 선호는 없습니다. 따라서 틀립니다. - B. 부모의 진술과 임상소견이 일치한다. 증후군 진단 시 부모의 진술이 의심스러운 경우가 많고, 임상소견과 일치하지 않을 수 있습니다. 따라서 틀립니다. - C. 신체 전 부위에 장기간에 걸친 타박상을 보인다. 매맞는아이증후군의 핵심 특징은 반복적 신체적 폭행으로 인한 타박상, 골절, 내장 손상 등 전신 손상입니다. "장기간"이라는 표현은 반복적 폭행의 과정을 반영하며, 이는 올바른 설명입니다. - D. 격렬하게 치아를 꽉 깨물어 구치부 치관파절이 발생된다. "치관파절"은 외상(예: 치아 충격)으로 발생하며, 폭행과 직접적 연관이 없습니다. 따라서 틀립니다. - E. 구강 내 손상의 특징은 하순 소대 주위의 점막열창과 치은파열이다. 구강 손상은 일반적으로 혀나 입술의 상처가 더 흔하며, 하순의 점막열창은 매맞는아이증후군의 특징이 아닙니다. 따라서 틀립니다. 결론: C는 매맞는아이증후군의 주요 임상적 특징을 정확히 반영합니다.
MAX	<think> 매맞는아이증후군은 부정적인 신체적 학대로 인해 발생하는 질환으로, 주로 1세에서 5세 사이의 어린이에서 흔하며, 특히 1세에서 3세 사이의 남자아이가 가장 많이 해당된다. 문제에서 제시된 보기 중, '8세 이하의 여자아이에게 호발한다(A)'는 사실과 일치하지 않으며, '부모의 진술과 임상소견이 일치한다(B)'는 상식적으로 틀린 표현이다. '격렬하게 치아를 꽉 깨물어 구치부 치관파절이 발생된다(D)'는 발음 과정 중의 상해로 인한 것이므로 매맞는아이증후군과는 관련이 없다. '구강 내 손상의 특징은 하순 소대 주위의 점막열창과 치은파열이다(E)'는 구타와는 상관이 없어 배제된다. '신체 전 부위에 장기간에 걸친 타박상을 보인다(C)'는 매맞는아이증후군의 대표적 임상 소견과 일치하므로, 정답은 'C'다. </think> 매맞는아이증후군(어린이 구타증)은 반복적인 신체적 학대로 인해 발생하는 질환으로, 주요 임상적 특징은 다음과 같습니다: **정답: C** **C. 신체 전 부위에 장기간에 걸친 타박상을 보인다.** **해설** 1. **A. 8세 이하의 여자아이에게 호발한다.** - 매맞는아이증후군은 **1세에서 3세 남자아이**에서 가장 흔합니다. 따라서 여성아이가 주된 대상은 아니며, 이는 **틀립니다**. 2. **B. 부모의 진술과 임상소견이 일치한다.** - 부모의 진술은 종종 학대와 관련이 있지만, 임상 소견(골절, 타박상 등)과 일치하지 않을 수 있습니다. 오히려 **일치하지 않는 경우가 많아** 이는 **틀립니다**. 3. **C. 신체 전 부위에 장기간에 걸친 타박상을 보인다.** - 반복적인 학대로 인해 **다양한 부위(골반, 뇌, 뼈 등)에 만성적 손상**이 관찰됩니다. 이는 매맞는아이증후군의 주요 임상적 특징이며, **정답입니다**. 4. **D. 격렬하게 치아를 꽉 깨물어 구치부 치관파절이 발생된다.** - 이는 **발음 과정 중의 상해**로, 학대와 직접적인 연관이 없습니다. **틀립니다**. 5. **E. 구강 내 손상의 특징은 하순 소대 주위의 점막열창과 치은파열이다.** - 구강 내 손상은 학대와 관련될 수 있지만, **이 특정 증상은 반드시 매맞는아이증후군의 필수 증상은 아닙니다**. **틀립니다**. **결론** 매맞는아이증후군은 **반복적 학대로 인한 다발성 타박상과 골절**이 특징입니다. 따라서 **C**가 정답입니다.

Table 9: A case study of ControlMed's reasoning under different reasoning length control modes (SHORT, MEDIUM, MAX) for a Korean clinical multiple-choice question about battered child syndrome. Across all modes, the model correctly identifies option C as the answer.