Hippocrates: An Open-Source Framework for Advancing Large Language Models in Healthcare

Emre Can Acikgoz[2], Osman Batur İnce[2], Rayene Bench[3], Arda Anıl Boz[4], Erkut Erdem[1], Aykut Erdem[2]

[1]Hacettepe University, [2]KUIS AI Center, Koç University [3]Yıldız Technical University [4]Robert College

Website

Motivation

We present **Hippocrates**, an open-source LLM framework specifically developed for the medical domain.

- In stark contrast to previous e□orts, it o□ers unrestricted access to its training datasets, codebase, checkpoints, and evaluation protocols.
- Also, we introduce Hippo, a family of 7B models tailored for the medical domain, fine-tuned from Mistral and LLaMA2 through continual pre-training, instruction tuning, and reinforcement learning from human and AI feedback.

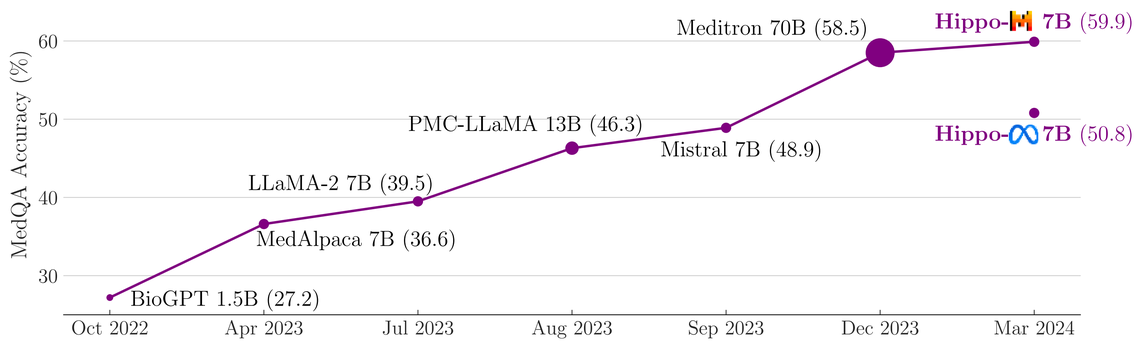

Figure 1: The evolution of medical LLM performances on the MedQA dataset. Our 7B Hippo- and Hippo- models achieve 50.8% and 59.9% 5-shot accuracy, respectively. Hippo- outperforms all existing open models, including even those with 70B parameters.

Hippocrates Framework

Hippocrates framework starts from domain-specific pre-training and progresses through supervised fine-tuning and reinforcement learning from AI-generated feedback to an extensive evaluation phase. This pipeline ensures our models are precisely tailored and rigorously tested for the medical domain.

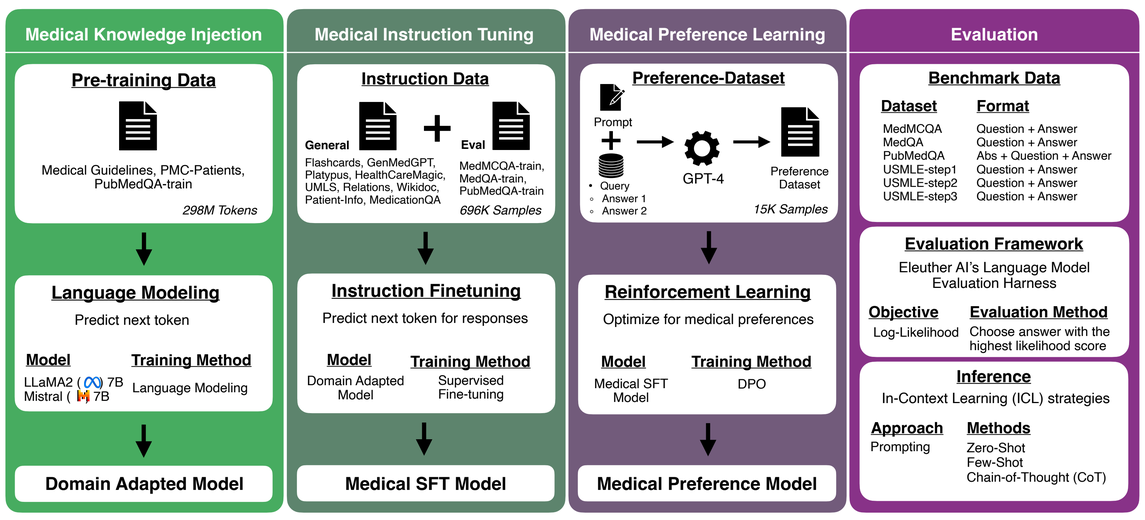

Figure 2: An overview of the Hippocrates framework, illustrating the four critical phases including (1) continued pre-training, (2) supervised fine-tuning, (3) reinforcement learning from AI-generated feedback, and (4) the comprehensive evaluation pipeline.

Experimental Setup

For an objective evaluation of domain-specific knowledge and reasoning capabilities in LLMs, a detailed and fair evaluation framework is essential. We selected **six widely recognized medical question-answering datasets**, namely MedMCQA, MedQA, PubMedQA, and USMLE Step 1-3. Performance metrics were derived through the use of the **EleutherAI evaluation framework**, ensuring a standardized approach to measuring model e□ectiveness in handling domain-specific queries.

Dataset	Source	Format	# Samples	# Choices	License
MedMCQA-test	MedMCQA	Question + Answer	4,183	4	MIT
MedQA-test	MedQA	Question + Answer	1,273	5	MIT
PubMedQA-test	PubMedQA	Abstract + Question + Answer	1,000	3	MIT
USMLE-step1	USMLE	Question + Answer	94	5	MIT
USMLE-step2	USMLE	Question + Answer	109	6	MIT
USMLE-step3	USMLE	Question + Answer	122	5	MIT

Results

We present a comparative analysis of our novel models, Hippo- and Hippo-, against a set of established base LLMs and medical-specific LLMs. Our evaluation includes both zero-shot and 5-shot learning scenarios. Hippo- and Hippo- not only **beat models with 7 billion and 13 billion parameters** but also **exceed the capabilities of those with 70 billion parameters**.

Model	MedMCQA 0-shot/5-shot	MedQA 0-shot/5-shot	PubmedQA 0-shot/5-shot	USMLE-1 0-shot/5-shot	USMLE-2 0-shot/5-shot	USMLE-3 0-shot/5-shot	Avg. 0-shot/5-shot
Gemma 2b	26.2/27.7	27.8/30.6	59.1/60.8	20.2/16.0	18.4/30.3	24.6/20.5	29.4/31.0
LLaMA-2 7b	34.4/39.4	29.3/39.5	72.3/72.4	18.1/22.3	22.9/33.0	27.1/32.0	34.0/39.8
Falcon 7b	30.5/31.8	27.9/31.0	65.3/64.4	18.1/25.5	26.6/20.2	23.8/25.4	32.0/33.0
Vicuna 7b	35.9/39.0	35.1/41.2	70.9/74.5	25.5/31.9	27.5/31.2	33.6/35.3	38.1/42.2
Mistral 7b	39.3/48.5	36.8/48.9	76.3/77.8	24.5/50.0	31.2/42.2	27.9/43.4	39.3/51.8
BioMedLM	32.2/29.6	29.3/30.6	55.2/55.2	15.9/22.3	19.3/18.4	23.0/31.2	25.9/31.2
BioGPT-Large	33.1/30.1	31.3/27.2	60.1/47.7	22.3/19.2	22.0/14.7	23.0/23.0	32.0/27.0
MedAlpaca 7b	35.8/37.5	36.1/36.6	73.2/70.6	22.3/27.7	27.5/32.1	29.5/37.7	37.4/40.4
PMC-LLaMA 7b	31.5/33.0	28.0/29.5	66.5/68.4	21.3/19.2	23.9/19.3	22.1/22.1	32.2/31.9
Meditron 7b	34.0/38.2	32.0/39.3	71.6/75.7	16.0/29.8	25.7/30.3	23.8/32.0	33.9/40.9
Bio-Mistral 7b	36.4/42.4	35.0/42.1	73.4/75.1	24.5/28.7	27.5/34.9	27.9/44.3	37.5/31.9
LLaMA-2 13b	38.2/43.9	34.3/43.3	75.9/71.9	20.2/38.3	22.0/29.4	23.0/38.5	35.6/40.9
Vicuna 13b	39.7/44.3	35.9/45.9	75.6/75.0	24.5/40.4	26.6/35.8	23.8/46.7	37.7/44.6
MedAlpaca 13b	32.5/33.3	31.8/34.3	72.6/72.5	24.5/23.4	24.5/26.6	30.3/29.5	36.0/44.2
PMC-LLaMA 13b	39.1/44.5	37.8/46.3	76.8/76.5	30.9/35.1	22.9/36.7	26.2/29.5	39.0/44.8
LLaMA-2 70b	42.8/52.0	44.9/56.1	73.2/77.8	31.9/59.6	44.0/57.8	44.3/53.3	46.8/59.4
Qwen 72b	50.5/59.2	47.7/53.4	77.2/76.8	45.7/67.0	43.1/56.9	38.5/61.5	50.5/62.5
ClinicalCamel 70b	43.7/53.4	45.5/58.5	73.6/77.6	40.4/59.6	43.1/60.6	42.6/60.7	48.2/61.7
Meditron 70b	43.4/51.9	44.9/58.5	76.4/**80.0**	35.1/57.5	41.3/56.9	37.7/59.8	46.5/60.8
Hippo- 7b	**54.3/53.9**	50.6/50.8	74.7/76.6	46.8/40.4	41.3/39.5	50.0/43.4	53.0/50.8
Hippo- 7b	49.7/51.8	**59.2**/59.9	**77.1**/78.1	**60.6/61.7**	**66.1/64.2**	**56.6**/56.6	**61.6/62.1**

Contribution of Each Training Stage

Hippo- and Hippo- analysis of Continued Pretraining, Instruction Tuning, and Direct Preference Optimization. The table **demonstrates the incremental impact** of Continued Pretraining (CP) on medical text data, Instruction Tuning (SFT), Direct Preference Optimization (DPO), and Chain-of-Thought (CoT) on the zero-shot capabilities of the LLaMA2 7B and Mistral 7B models across a range of medical benchmarks.

Model	MedMCQA	MedQA	PubmedQA	USMLE-1	USMLE-2	USMLE-3	Avg.
LLaMA2 7b	34.4	29.3	72.3	18.1	22.9	27.1	34.0
+ CP	34.6	31.9	72.8	20.2	25.7	21.3	34.4
+ SFT	52.7	49.7	**75.7**	37.2	42.2	44.3	50.3
+ CP + SFT	54.3	**50.6**	74.7	46.8	41.3	**50.0**	**53.0**
+ CP + SFT + DPO	**54.4**	50.4	74.8	46.8	39.5	49.2	52.5
+ CP + SFT + DPO + CoT	54.0	50.3	73.3	**48.9**	**43.7**	45.1	52.6
Mistral 7b	39.3	36.8	76.3	24.5	31.2	27.9	39.3
+ CP	40.5	37.2	74.9	29.8	33.9	29.5	41.0
+ SFT	49.7	59.2	77.1	**60.6**	**66.1**	56.6	**61.6**
+ CP + SFT	**51.5**	**60.9**	76.5	55.3	65.1	57.4	61.1
+ CP + SFT + DPO	49.3	57.3	**77.3**	56.4	62.4	54.9	59.6
+ CP + SFT + DPO + CoT	51.0	**60.9**	63.5	59.6	59.6	**63.9**	59.8

Uncertainty Quantification

We conducted an uncertainty quantification experiment on Hippo- to understand its performance on the evaluation datasets. Our findings reveal that:

- Our model assigns higher probabilities to questions it answers correctly across all datasets, suggesting an ability to self-calibrate its certainty.
- The model's confidence is notably higher on MedMCQA and lower on PubMedQA, possibly reflecting the datasets' relative simplicity and complexity, respectively.
- Additionally, the model's confidence changes with di□erent training stages.

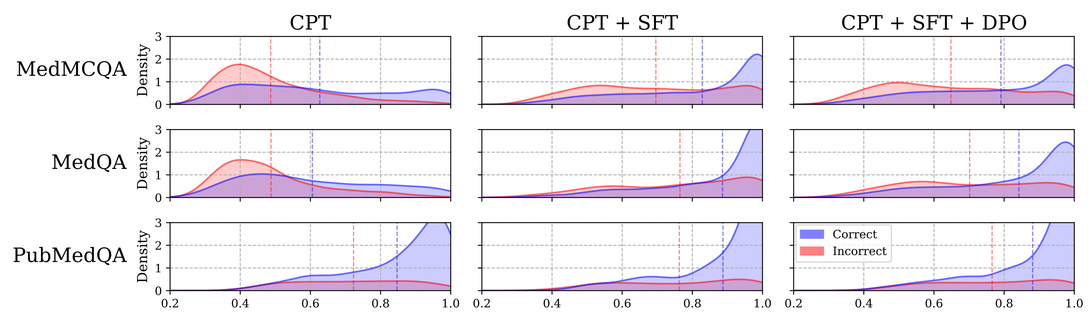

Figure 3: Uncertainty quantification for our best-performing 5-shot Hippo- model., where we plot the probability distributions assigned by the model to both correct predictions and incorrect predictions on the MedMCQA, MedQA, and PubMedQA datasets.