清華大學
Tsinghua University

UltraMedical: Building Specialized Generalists in Biomedicine

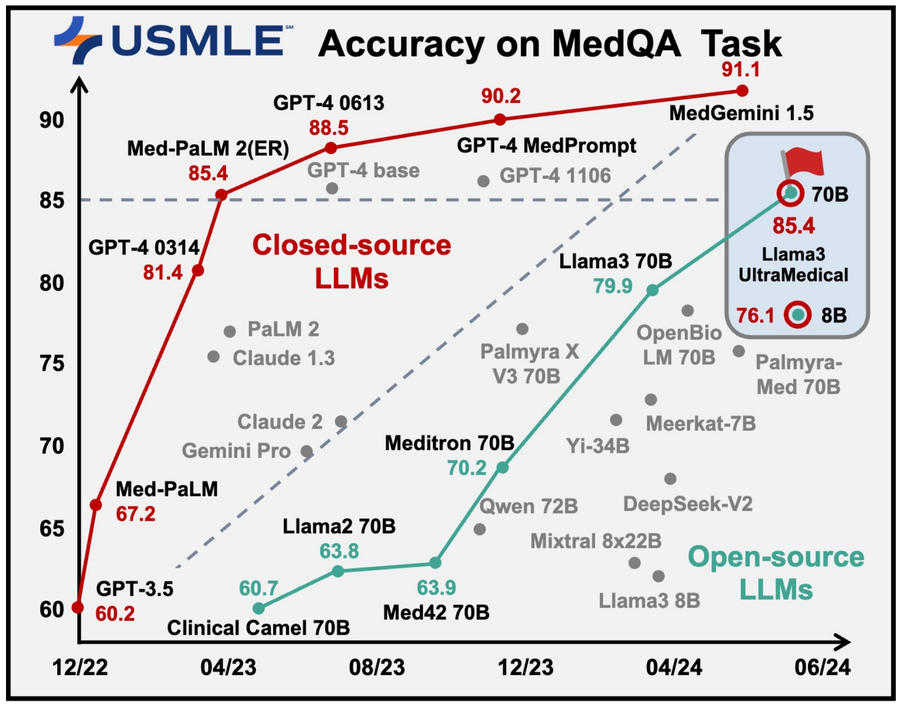
Kaiyan Zhang[α,ε] **Sihang Zeng**[β] **Ermo Hua**[α,ε] **Ning Ding**[α*] **Zhang-Ren Chen**[γ]
Zhiyuan Ma[α] **Haoxin Li**[α] **Ganqu Cui**[α] **Biqing Qi**[α] **Xuekai Zhu**[δ] **Xingtai Lv**[α,ε]
Jin-Fang Hu[γ] **Zhiyuan Liu**[α] **Bowen Zhou**[α*]

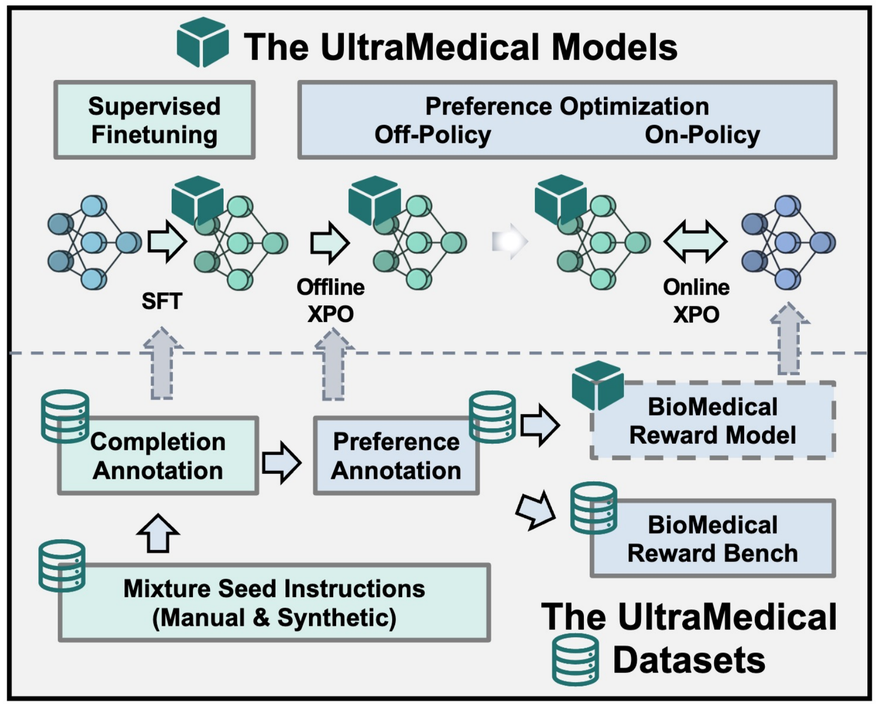
[α] Tsinghua University [β] University of Washington
[γ] The First Affiliated Hospital of Nanchang University
[δ] Shanghai Jiao Tong University [ε] Frontis.AI

Summary

- We present the UltraMedical collections, which consist of high-quality manual and synthetic datasets in the biomedicine domain, featuring preference annotations across multiple advanced LLMs.

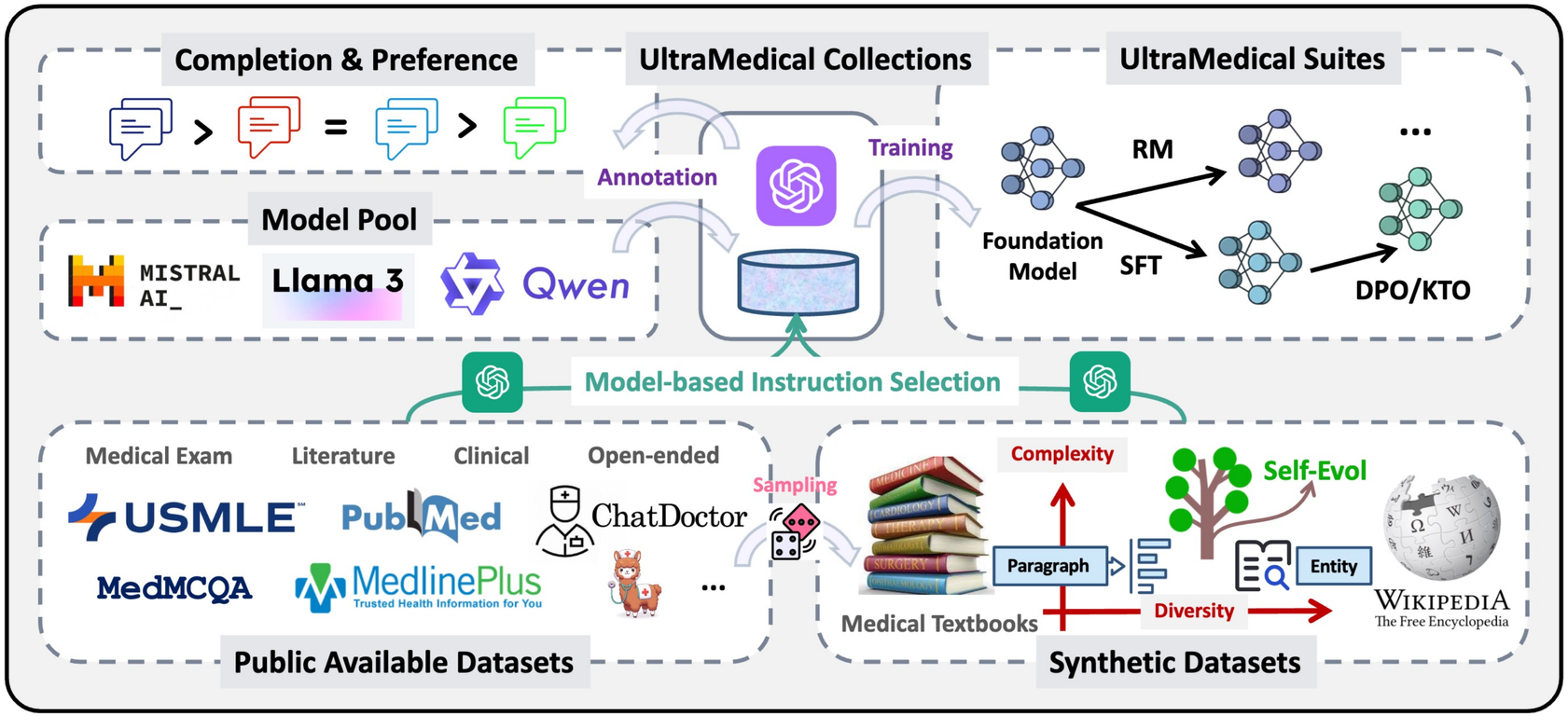
- Our 8B model significantly outperforms previous larger models such as MedPaLM 1, Gemini-1.0, GPT-3.5, and Meditron-70B. Moreover, our 70B model achieved an 86.5 on MedQA-USMLE, marking the highest result among open-source LLMs and comparable to MedPaLM 2 and GPT-4.





UltraMedical: Dataset

- Instruction Composition
  - Principle of Diversity: medical exam, literature, clinical and research questions
  - Principle of Complexity: model-based ranking and two step self-evolution



UltraMedical: Dataset

- Synthetic Dataset: MedQA-Evol, TextBookQA, and WikiInstruct
- Completion Annotation: gpt-4-turbo with chain-of-thought prompting

Table 1: Instructions Statistics. Datasets marked with "★" represent our customized synthetic data, while the others are adapted from publicly available data. Average length and score by ChatGPT noted as *Avg.Len* and *Avg.Score*.

Category	Synthetic	Dataset	# Original	Avg.Len	Avg.Score	# Retained
Examination	✗	MedQA	10.2K	128.94	7.35	9.3K
	✗	MedMCQA	183K	23.12	4.73	59K
	✓	★ MedQA-Evol	51.8K	76.52	8.07	51.8K
	✓	★ TextBookQA	91.7K	75.92	7.72	91.7K
Literature	✗	PubMedQA	211K	218.2	7.95	88.7K
Open-ended	✗	ChatDoctor	100K	98.93	6.83	31.1K
	✗	MedQuad	47K	8.21	4.54	6K
	✓	MedInstruct-52K	52K	36.05	5.25	23K
	✓	MedIns-120K	120K	84.93	5.36	25K
	✓	★ WikiInstruct	23K	46.73	8.8	23K
★ **UltraMedical (Mixed)**		**Instructions**	-	101.63	8.2	**410K**
		Preference Pairs	**1.8M**	-	-	**100K**

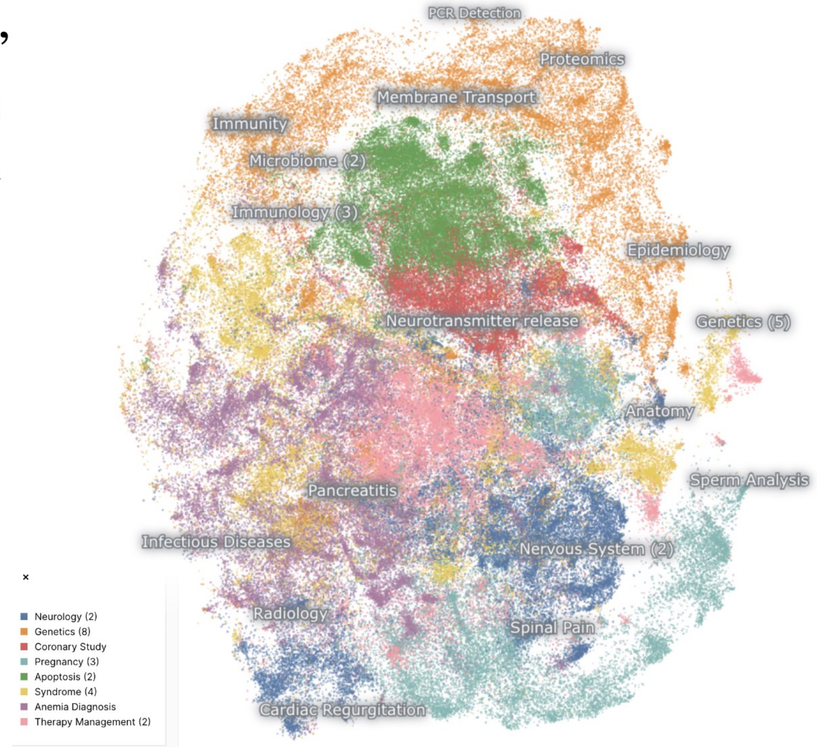
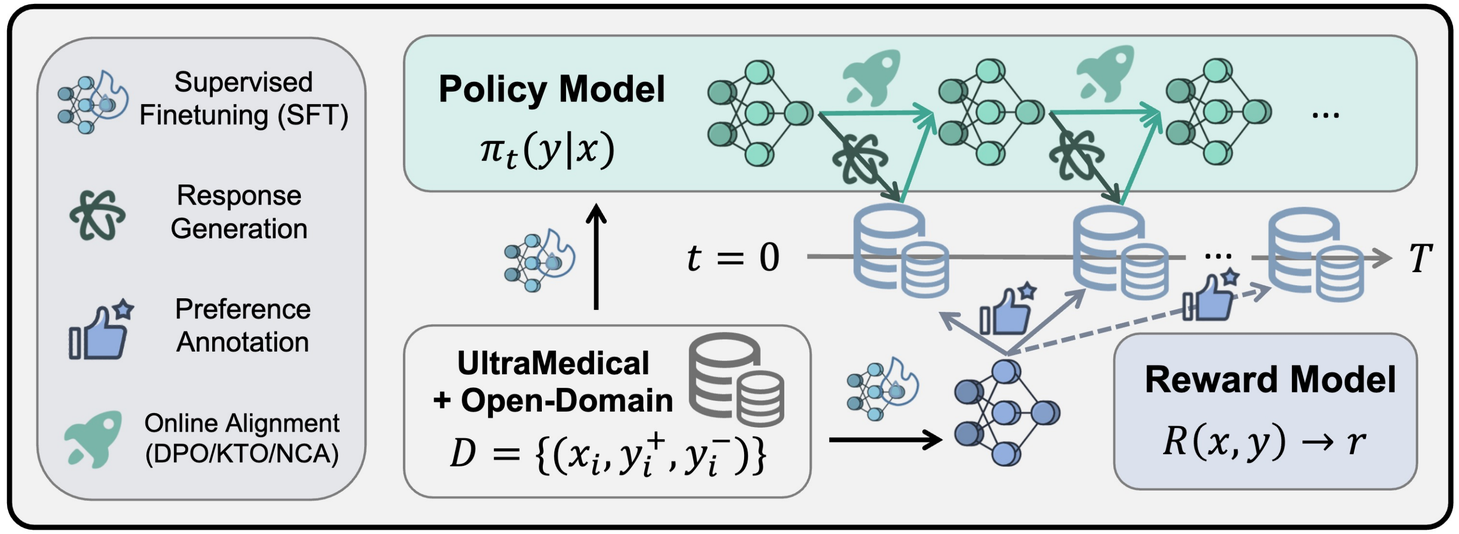


Figure 3: Broad Topics Distribution

UltraMedical: Models

- **Step 1: Supervised Fine-tuning.**
  - 410K medical domain and 190K open-domain samples (gpt-4-turbo)
- **Step 2: Preference Learning.**
  - 100K medical domain and 75K open-domain pairs
- **Step 3: Reward Modeling.**
  - Training Outcome-level reward model with UltraSeries
- **Step 4: Iterative Preference Learning.**
  - Best-of-N on-policy sampling with K times



UltraMedical: Results

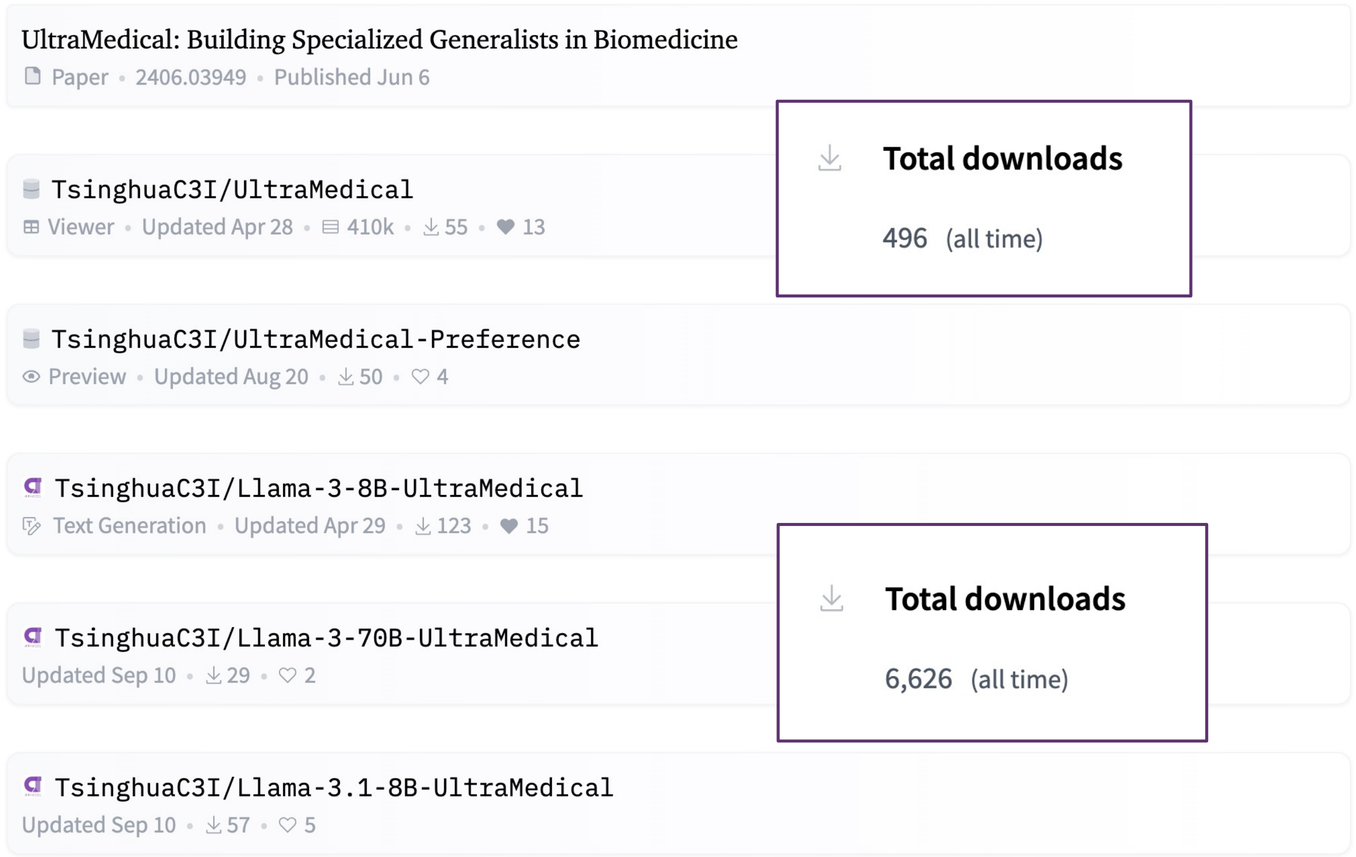
- Data mixture of both medical and general enhances both SFT and xPO
- Online preference learning enhance performance than offline learning
- There is a trade-off performance in medical and open domain

Instruct Model	K-QA		MT-Bench	AlpacaEval 2		MMLU	GPQA	GSM8K
	Compl. (↑)	Hall. (↓)	GPT-4	LC (%)	WR (%)	5-shot	0-shot	8-shot, CoT
Mistral-7B-Instruct	0.5335	0.2090	6.84	17.1	14.7	58.4	26.3	39.9
Llama-3-8B-Instruct	0.6037	0.1940	8.10	22.9	22.6	68.4	34.2	79.6
OpenBioLM-8B	0.3135	0.1194	4.38	0.06	0.25	44.2	24.8	41.6
★ UltraMedLM 8B	0.7242	0.0945	7.64	30.7	31.9	68.1	34.2	75.9
Mixtral-8x7B	0.6617	0.1343	8.30	23.7	18.3	70.6	39.5	93.0
Llama-3-70B-Instruct	0.6545	0.1357	9.01	34.4	33.2	82.0	39.5	93.0
OpenBioLM-70B	0.5951	0.1100	8.53	30.8	31.0	60.1	29.2	90.5
★ UltraMedLM 70B	0.6077	0.0896	8.54	33.0	32.1	77.2	39.7	88.7
GPT-3.5-Turbo (1106)	0.6208	0.0746	8.32	19.3	9.2	70.0	28.1	57.1
GPT-4-Turbo (1106)	0.6390	0.1095	9.32	50.0	50.0	86.4	49.1	92.0

Instruct Model & Task	MedQA (US 4-opt)	MedMCQA (Dev)	PubMedQA (Reasoning)	MMLU						Avg.
				Clinical knowledge	Medical genetics	Anatomy	Professional medicine	College biology	College medicine	
~7B Models (0-shot CoT)										
Mistral-7B-Instruct*	37.0	31.9	44.2	51.7	57.0	51.1	47.4	42.2	43.4	45.10
Starling-LM-7B-beta*	50.6	45.3	67.2	66.4	67.0	57.8	64.0	67.4	60.7	60.71
BioMistral-7B	46.6	45.7	68.1	63.1	63.3	49.9	57.4	63.4	57.8	57.26
Meerkat-7B (Ens)	**74.3**	**60.7**	-	61.9	70.4	61.5	69.5	55.4	57.8	63.94
Llama-3-8B-Instruct*	60.9	50.7	73.0	72.1	76.0	63.0	77.2	79.9	64.2	68.56
Internist-7B	60.5	55.8	**79.4**	70.6	71.0	65.9	76.1	-	63.0	67.79
OpenBioLLM-8B	59.0	56.9	74.1	**76.1**	**86.1**	**69.8**	**78.2**	**84.2**	**68.0**	**72.48**
★ *Llama-3-8B UltraMedical (Our)*										
UltraMed + SFT	73.3	61.5	77.0	78.9	78.0	74.1	83.8	78.5	71.7	75.20
UltraMed + Vanilla DPO	73.7	63.6	78.2	76.2	88.0	75.6	83.8	79.9	70.5	76.61
UltraMed + Vanilla KTO	72.7	63.3	79.2	77.0	87.0	69.6	86.4	81.9	72.3	76.61
UltraMix + SFT	74.5	62.0	79.2	75.8	83.0	73.3	83.5	81.2	70.5	75.90
UltraMix + Vanilla DPO	74.9	63.6	79.4	78.1	84.0	71.9	86.8	80.6	76.3	77.29
UltraMix + Vanilla KTO	73.3	63.8	79.0	77.4	87.0	71.9	85.3	80.6	72.3	76.74
UltraMix + Iterative DPO	74.2	62.7	79.2	78.1	87.0	76.3	87.5	82.6	69.9	77.51
UltraMix + Iterative KTO	74.8	63.6	78.8	77.0	91.0	75.6	83.8	79.9	72.3	77.41
UltraMix Best (Ens)	**76.1**	**65.3**	**79.0**	**77.7**	**87.0**	**74.8**	**87.1**	**82.6**	**75.1**	**78.32**
>40B Models (0-shot CoT)										
Med42-70B	66.6	60.6	67.2	76.6	77.0	66.7	79.8	75.7	66.5	70.74
Mixtral-8x7B-Instruct*	52.8	49.7	46.2	71.7	70.0	62.2	71.0	77.8	67.1	63.17
Mixtral-8x22B-Instruct*	73.1	63.3	71.4	84.2	89.0	77.0	88.2	88.2	78.0	79.16
Qwen1.5-72B-Chat*	63.6	59.0	32.4	78.9	80.0	68.9	82.7	91.0	75.7	70.24
Llama-2-70B-Chat*	47.3	41.9	63.8	64.9	70.0	54.1	59.2	66.7	61.3	58.80
Llama-3-70B-Instruct*	**79.9**	69.6	75.8	87.2	93.0	76.3	88.2	92.4	81.5	82.66
DeepSeek-v2-Chat*	68.6	61.5	71.0	83.0	90.0	73.3	86.8	88.9	78.0	77.90
OpenBioLLM-70B	78.2	**74.0**	**79.0**	92.9	93.2	83.9	**93.8**	93.8	**85.7**	**86.06**
OpenBioLLM-70B (Ens)*	77.5	73.7	**79.0**	**93.6**	**95.0**	**85.9**	87.9	**95.1**	85.5	85.92
★ *Llama-3-70B UltraMedical (Our)*										
UltraMed + SFT	82.2	72.3	78.8	86.4	91.0	82.2	92.3	89.6	86.7	84.62
UltraMed + Vanilla DPO	85.3	73.0	78.8	86.4	92.0	84.4	94.1	91.7	84.4	85.57
UltraMed + Vanilla KTO	84.7	73.0	79.8	86.0	93.0	84.4	92.6	93.1	81.5	85.35
UltraMix + SFT	83.7	73.0	77.6	84.9	94.9	80.7	91.9	91.0	81.5	84.27
UltraMix + Vanilla DPO	84.0	74.1	77.4	85.7	95.0	80.7	93.8	94.4	85.0	85.56
UltraMix + Vanilla KTO	84.8	73.2	80.0	86.8	92.0	84.4	93.8	93.1	84.4	85.84
UltraMix Best (Ens)	**85.4**	**74.7**	**78.8**	**89.4**	**95.0**	**85.2**	**92.6**	**95.1**	**82.1**	**86.49**
Proprietary Models (Mixed - few-shot, self-consistency)										
GPT-3.5-Trubo	57.7	72.7	53.8	74.7	74.0	65.9	72.8	72.9	64.7	67.70
Flan-PaLM (best)	67.6	57.6	79.0	80.4	75.0	63.7	83.8	88.9	76.3	74.70
GPT-4 (5-shot)	81.4	72.4	75.2	86.4	92.0	80.0	93.8	95.1	76.9	83.69
GPT-4 (0-shot CoT)	85.8	72.3	70.0	90.2	94	84.4	94.5	93.8	83.2	85.36
Med-PaLM 2 (ER)	85.4	72.3	75.0	88.7	92.0	84.4	92.3	95.8	83.2	85.46
GPT-4-base (5-shot)	86.1	73.7	80.4	88.7	97.0	85.2	93.8	97.2	80.9	87.00
GPT-4 (Medprompt)	**90.2**	**79.1**	**82.0**	**95.8**	**98.0**	**89.6**	**95.2**	**97.9**	**89.0**	**90.76**

Open Source

- All the models and datasets are released on **Huggingface and GitHub**
- The total downloads of models and datasets are more than **7,000 and 6,00 times**, respectively



https://huggingface.co/TsinghuaC3I

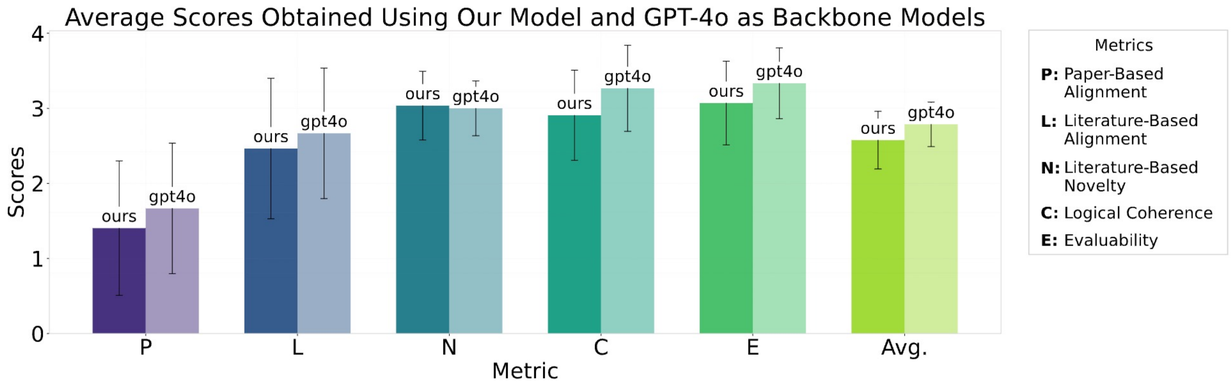
Hugging Face

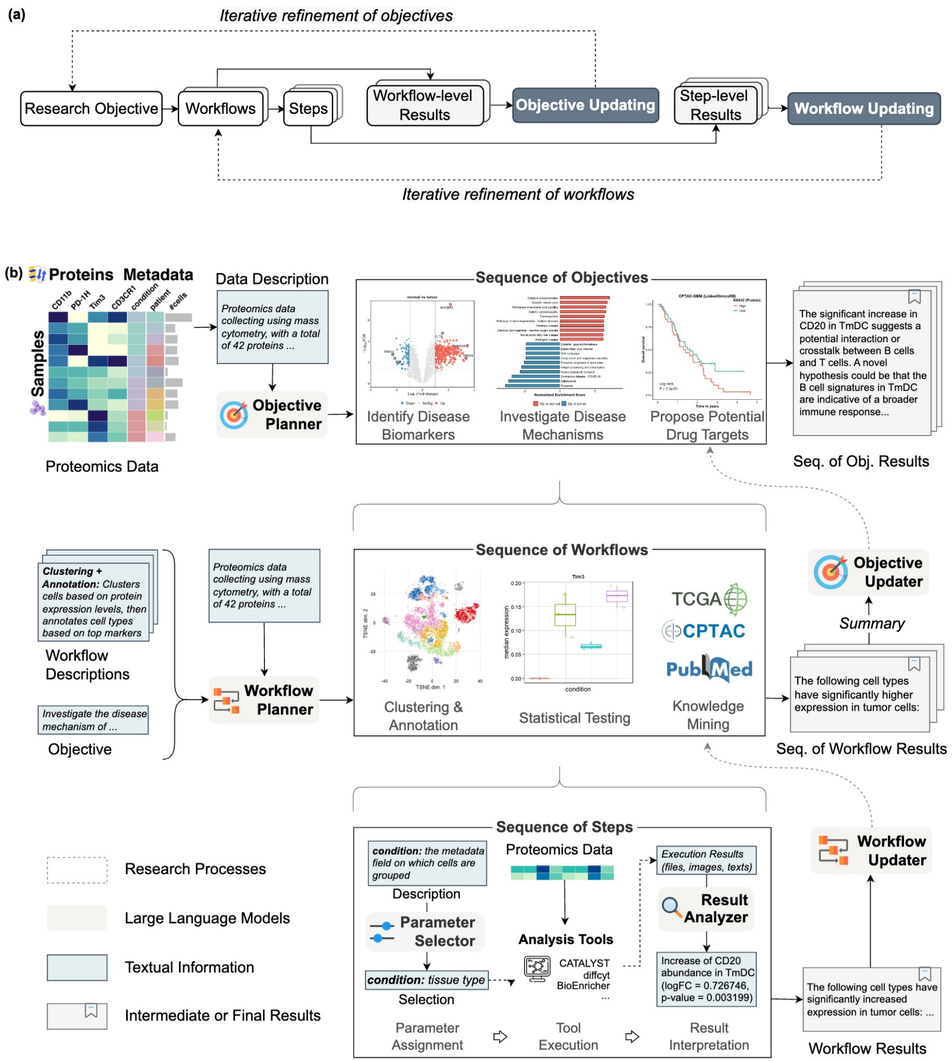
https://github.com/TsinghuaC3I/UltraMedical

GitHub

UltraMedical for Scientific Discovery

- Application
  - Building fully automated system for scientific discovery from raw proteomics data
- Performance
  - UltraMedical demonstrates competitive performance compared to the state-of-the-art gpt-4o models





Ding, Ning, Shang Qu, Linhai Xie, Yifei Li, Zaoqu Liu, Kaiyan Zhang, Yibai Xiong et al. "Automating Exploratory Proteomics Research via Language Models." *arXiv preprint arXiv:2411.03743* (2024).

Thanks