Benchmarking Large Language Models on **CMExam** - A Comprehensive **C**hinese **M**edical **Exam** Dataset

Junling Liu*†, Peilin Zhou*, Yining Hua*#, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, Michael Lingzhi Li
*Equal contribution; †Correspondence; #Presenter

CMExam (dataset)

- **Large** (60k+ question-answer pairs)
- **Reliable** (Chinese National Medical Licensing Examination and other official resources)
- **Quantitative** (Five additional annotations)
- **Qualitative** (Corresponding explanation)

Benchmark

- **High-coverage** (20 models)
- **Multifaceted** (Answer prediction + reasoning)

Challenge of evaluating LLMs in medical fields

- Insufficient size and diversity
- Lack clear choice evaluations
- Lack explanations
- Unreliable sources
- Language resource inequality

Table 1: A review of medical QA datasets. * indicates availability of additional annotations with authoritative references, † indicates availability of benchmarks, and ‡ indicates datasets with more than 50K questions

Language	Data Source Type	Question Type	
		Multiple Choice	Open-ended
English	Consumer Questions	MedMCQA (Pal et al., 2022)	LiveQA-Med (Abacha et al., 2017) CliCR‡ (Šuster and Daelemans, 2018) HealthQA (Zhu et al., 2019) MEDIQA (Abacha et al., 2019b) emrQA‡ (Pampari et al., 2018) MedQuaD (Ben Abacha and Demner-Fushman, 2019) MedicationQA* (Abacha et al., 2019a) MEDIQA-AnS (Savery et al., 2020) MASH-QA (Zhu et al., 2020)
	Research, Books, or Exams	MEDQA‡(Jin et al., 2021) MMLU†‡ (Hendrycks et al., 2020) MedMCQA (Pal et al., 2022) MultiMedQA*† (Singhal et al., 2022)	BioASQ (Krithara et al., 2023) MultiMedQA*† (Singhal et al., 2022)
Chinese	Consumer Questions	-	webMedQA*‡ (He et al., 2019) cMedQA-v1.0‡ (Zhang et al., 2017) cMedQA-v2.0‡ (Zhang et al., 2018) ChiMed (Tian et al., 2019) Huatuo-26M†‡ (Li et al., 2023)
	Research, Books, or Exams	MLEC-QA‡ (Zeng et al., 2023a) **CMExam***†‡(ours)	MLEC-QA‡ (Zeng et al., 2023a) **CMExam***†‡(ours)

The CMExam Dataset

- Sourced from past exams and practice questions

- 60K+ QA pairs
- Five Additional Annotations
  1. Disease Groups
  2. Clinical Departments
  3. Medical Disciplines
  4. Areas of Competency
  5. Question Difficulty Levels
- Corresponding Explanation

Example data point

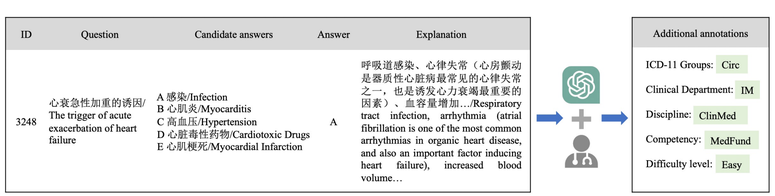

Figure 1: An example question of CMExam. Abbreviations: Circulatory System Diseases (Circ), Internal Medicine (IM), Clinical Medicine (ClinMed), Medical Fundamentals (MedFund).

Statistics of CMExam

Table 14: Basic statistics of CMExam. Q: questions; E: explanations; Q1/3: the first/ third quantile.

	Train	Dev	Test	Total
Question #	54,497	6,811	6,811	68,119
Vocab	4,545	3,620	3,599	4,629
Max Q tokens	676	500	585	676
Max E tokens	2,999	2,678	2,680	2,999
Avg Q tokens	29.78	30.07	32.63	30.83
Avg E tokens	186.24	188.95	201.44	192.21
Median (Q1, Q3) Q tokens	17 (12, 32)	18 (12, 32)	18 (12, 37)	18 (12, 32)
Median (Q1, Q3) E tokens	146 (69, 246)	143 (65, 247)	158 (80, 263)	146 (69, 247)

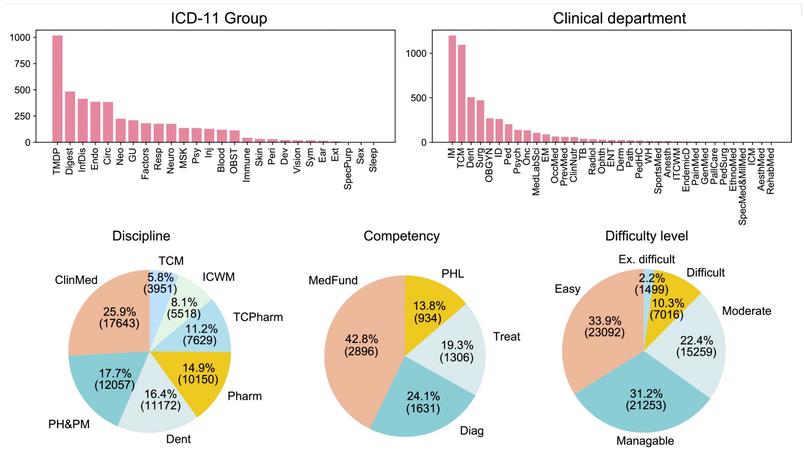

Figure 2: Additional CMExam statistics. For the question length distribution subplot, only the portion within IQR is shown.

Benchmark Results

Table 3: Overall comparison on CMExam dataset. We **bold** the best result and underline the second best result.

Model type	Models	size	Prediction		Reasoning				
			Acc (%)	F1 (%)	BLEU-1	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
General Domain	GPT-3.5-turbo	175B	46.4±0.6	46.1±0.7	3.56±0.67	1.49±0.51	33.80±0.19	16.39±0.18	14.83±0.13
	GPT-4	-	**61.6±0.1**	**61.7±0.1**	0.17±0.00	0.06±0.00	29.74±0.09	14.84±0.04	11.51±0.03
	ChatGLM	6B	26.3±0.0	25.7±0.1	16.51±0.08	5.00±0.06	35.18±0.11	15.73±0.05	17.09±0.13
	LLaMA	7B	0.4±0.0	0.3±0.0	11.99±0.03	5.70±0.0	27.33±0.06	11.88±0.03	10.78±0.04
	Vicuna	7B	5.0±0.0	4.8±0.1	20.15±0.01	9.26±0.01	38.43±0.02	16.90±0.01	16.33±0.01
	Alpaca	7B	8.5±0.0	8.4±0.0	4.75±0.00	2.50±0.00	22.52±0.00	9.54±0.00	8.40±0.00
Medical Domain	Huatuo	7B	12.9±0.0	7.0±0.0	0.21±0.00	0.12±0.00	25.11±0.08	11.56±0.04	9.73±0.02
	MedAlpaca	7B	20.0±0.0	10.7±0.0	0.00±0.00	0.00±0.00	1.90±0.00	0.04±0.00	0.52±0.03
	DoctorGLM	6B	-	-	9.43±0.09	2.65±0.03	21.11±0.03	6.86±0.01	9.99±0.06
	PromptCLUE-base-CMExam	0.1B	-	-	18.75±0.08	6.65±0.05	40.88±0.11	21.90±0.11	18.31±0.11
	Bart-base-chinese-CMExam	0.1B	-	-	23.00±0.40	10.35±0.16	44.33±0.09	24.29±0.09	20.80±0.09
	Bart-large-chinese-CMExam	0.1B	-	-	26.37±0.18	11.65±0.08	44.92±0.12	24.34±0.12	21.75±0.03
	BERT-CMExam	0.1B	31.8±0.2	31.2±0.2	-	-	-	-	-
	RoBERTa-CMExam	0.3B	37.1±0.1	36.7±0.4	-	-	-	-	-
	MedAlpaca-CMExam	7B	30.5±0.1	30.4±0.1	16.35±0.80	9.78±0.47	44.31±0.85	27.05±0.50	24.55±0.43
	Huatuo-CMExam	7B	28.6±0.5	29.3±0.2	29.04±0.01	16.72±0.03	43.85±0.24	25.36±0.22	21.72±0.24
	ChatGLM-CMExam	6B	45.3±1.4	45.2±1.4	**31.10±0.23**	**18.94±0.12**	43.94±0.28	**31.48±0.14**	**29.39±0.14**
	LLaMA-CMExam	7B	18.3±0.5	20.6±0.5	29.25±0.23	16.46±0.10	**45.88±0.04**	26.57±0.04	23.31±0.02
	Alpaca-CMExam	7B	21.1±0.6	24.9±0.4	29.57±0.10	16.40±0.12	45.48±0.12	25.53±0.18	22.97±0.06
	Vicuna-CMExam	7B	27.3±0.5	28.2±0.3	29.82±0.03	17.30±0.01	44.98±0.16	26.25±0.13	22.44±0.09
Random	Random	-	3.1±0.2	5.1±0.3	-	-	-	-	-
Human Performance	Human volunteers	-	71.6	-	-	-	-	-	-

Example: Performance Stratified by Difficulty

Table 8: Results by question difficulty.

Categories	GPT-4	GPT-3.5	ChatGLM	ChatGLM-CMExam	Average
Easy	74.6±0.1	58.5±0.6	31.4±0.2	61.5±0.3	56.5±0.4
Manageable	63.9±0.2	47.4±0.7	25.9±0.5	46.1±0.3	45.8±0.6
Moderate	51.3±0.6	36.8±0.8	23.0±0.4	34.5±0.6	36.4±0.7
Difficult	36.4±0.9	26.2±0.7	18.9±0.5	24.3±0.9	26.5±0.6
Extremely difficult	27.2±1.0	21.4±2.2	15.8±1.0	12.2±1.1	19.1±1.1

Quality of Model-generated Explanations

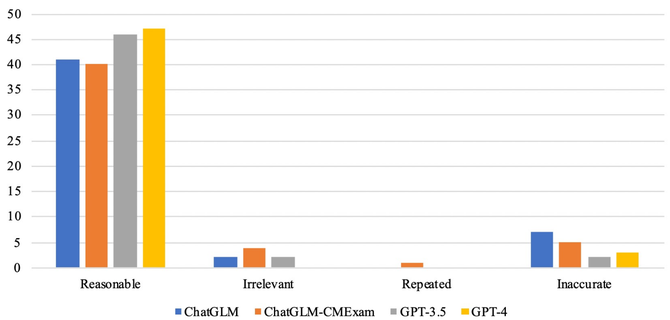

Figure 9: Correctness analysis.

Limitations

- Excluding non-textual questions may introduce biases.
- BLEU and ROUGE metrics are inadequate for fully assessing explanations; better expert analysis needed in future.

Ethics

- Adheres to legal and ethical guidelines.
- Authenticated and accurate for evaluating LLMs.
- Intended for academic/research use only; commercial misuse prohibited.
- Users should acknowledge dataset limitations and specific context.
- Not for assessing individual medical competence or patient diagnosis.

Future Directions

- Translate to English (in-progress)
- Include multimodal information (check out our new dataset ChiMed-VL-Instruction - 469,441 vision-language QA pairs, link in QR code)