KnowLab, Institute of Health Informatics

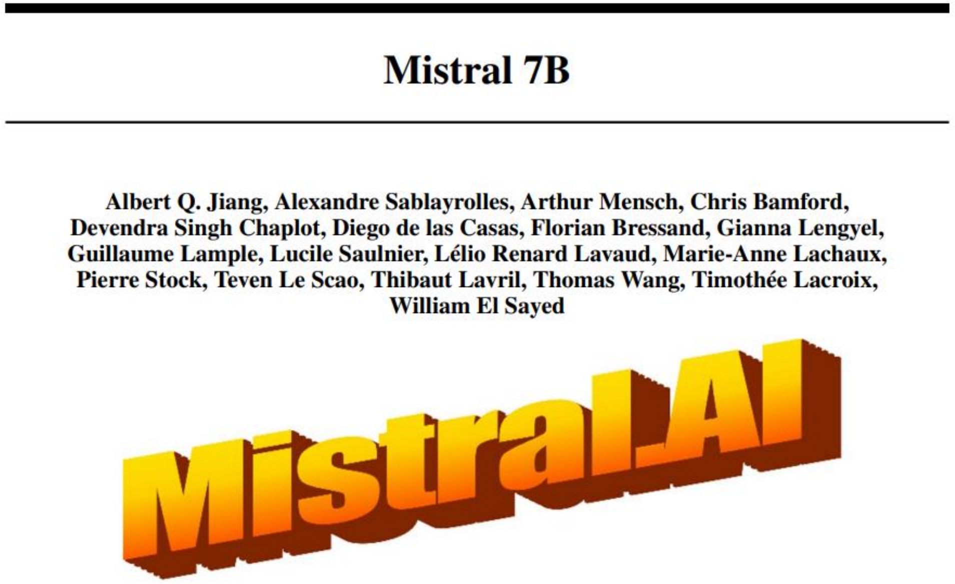
Mistral 7B and Medical Language Model Benchmark

Yunsoo Kim

I. Mistral 7B – Introduction

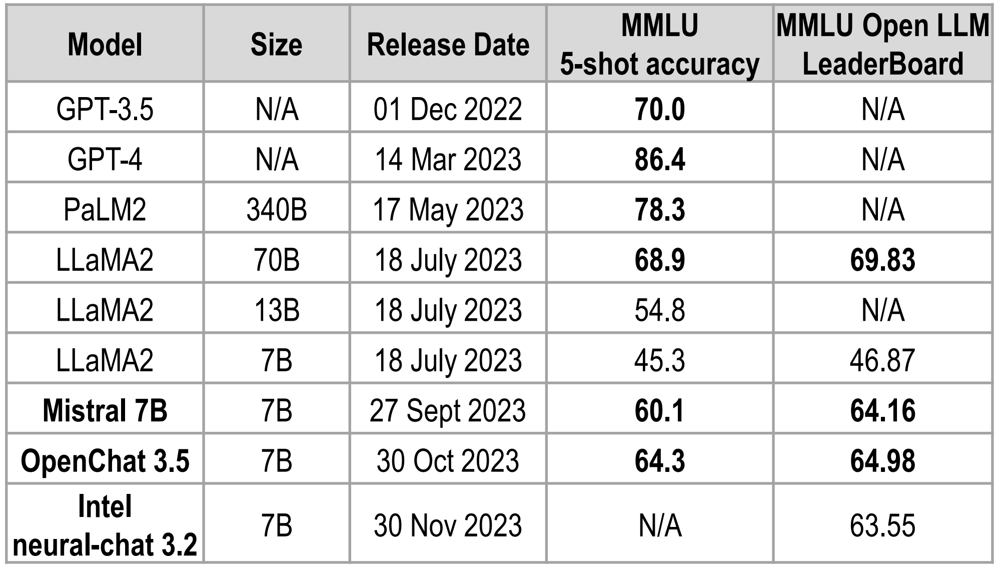
Mistral 7B was released 27 September 2023

- Outperformed the best open 13B model (LLaMA2) in all evaluated benchmarks including MMLU.
- Focused on the inference cost → "Obtain the best performance with the smallest possible model"
- Facilitating the small language model (**SLM**) research: new way of reinforcement learning



Mistral 7B

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed



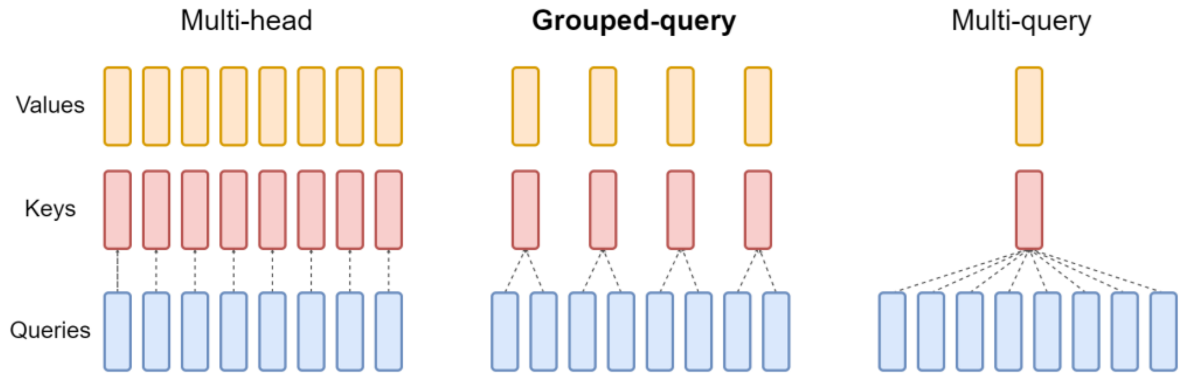
Model	Size	Release Date	MMLU 5-shot accuracy	MMLU Open LLM LeaderBoard
GPT-3.5	N/A	01 Dec 2022	**70.0**	N/A
GPT-4	N/A	14 Mar 2023	**86.4**	N/A
PaLM2	340B	17 May 2023	**78.3**	N/A
LLaMA2	70B	18 July 2023	**68.9**	**69.83**
LLaMA2	13B	18 July 2023	54.8	N/A
LLaMA2	7B	18 July 2023	45.3	46.87
Mistral 7B	7B	27 Sept 2023	**60.1**	**64.16**
OpenChat 3.5	7B	30 Oct 2023	**64.3**	**64.98**
Intel neural-chat 3.2	7B	30 Nov 2023	N/A	63.55

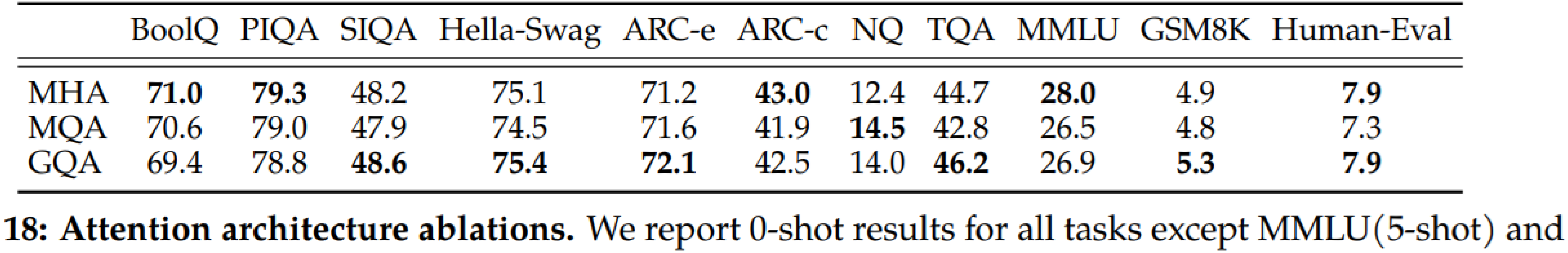
Jiang *et al.*, 2023 [2310.06825] Mistral 7B

I. Mistral 7B – Efficient Model (1)

Grouped Query Attention for the efficiency: Solves GPU Out-of-Memory (OOM) issue

- Released in May 2023 by Google
- Also used by LLaMA2 70B model
- From LLaMA2 – MHA OOM with a batch size of 128 for 2k context





	BoolQ	PIQA	SIQA	Hella-Swag	ARC-e	ARC-c	NQ	TQA	MMLU	GSM8K	Human-Eval
MHA	**71.0**	**79.3**	48.2	75.1	71.2	**43.0**	12.4	44.7	**28.0**	4.9	**7.9**
MQA	70.6	79.0	47.9	74.5	71.6	41.9	**14.5**	42.8	26.5	4.8	7.3
GQA	69.4	78.8	**48.6**	**75.4**	**72.1**	42.5	14.0	**46.2**	26.9	**5.3**	**7.9**

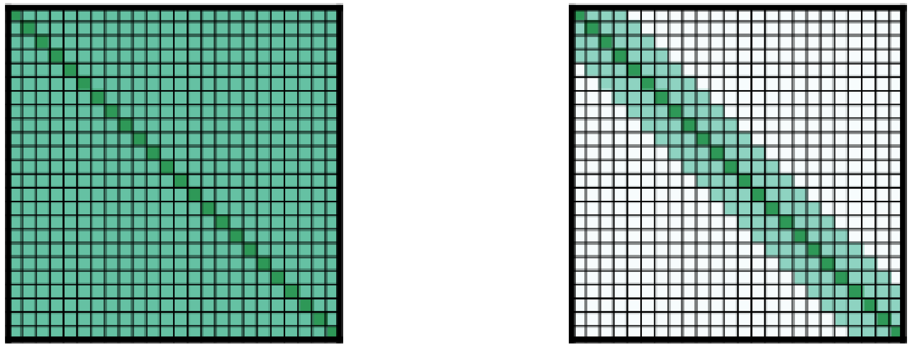
Table 18: Attention architecture ablations. We report 0-shot results for all tasks except MMLU(5-shot) and

Ainslie *et al.*, 2023 [2305.13245] GQA
Touvron *et al.*, 2023 [2307.09288] Llama 2: Open Foundation and Fine-Tuned Chat Models

I. Mistral 7B – Efficient Model (2)

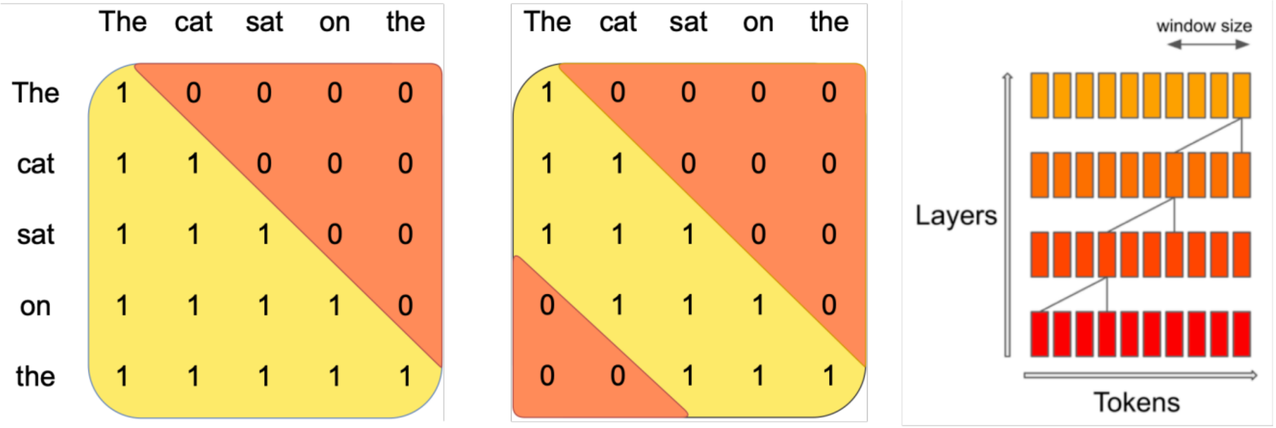
Sliding Window Attention for the efficiency: **2X faster** for sequence length, n=16k, and window size, w=4k

- Modified version of the sliding window attention from Longformer (Encoder or Encoder-Decoder Model)
- Similar to CNNs – uses windowed attention
- Self attention O(n^2) ➔ Sliding window attention O(n x w)
- With multiple layers, the last layer can build representations from the entire input



(a) Full n^2 attention

(b) Sliding window attention



	The	cat	sat	on	the
The	1	0	0	0	0
cat	1	1	0	0	0
sat	1	1	1	0	0
on	1	1	1	1	0
the	1	1	1	1	1

	The	cat	sat	on	the
The	1	0	0	0	0
cat	1	1	0	0	0
sat	1	1	1	0	0
on	0	1	1	1	0
the	0	0	1	1	1

Vanilla Attention **Sliding Window Attention** **Effective Context Length**

LongFormer : Both-sided window attends to $\frac{1}{2}$w tokens on each side

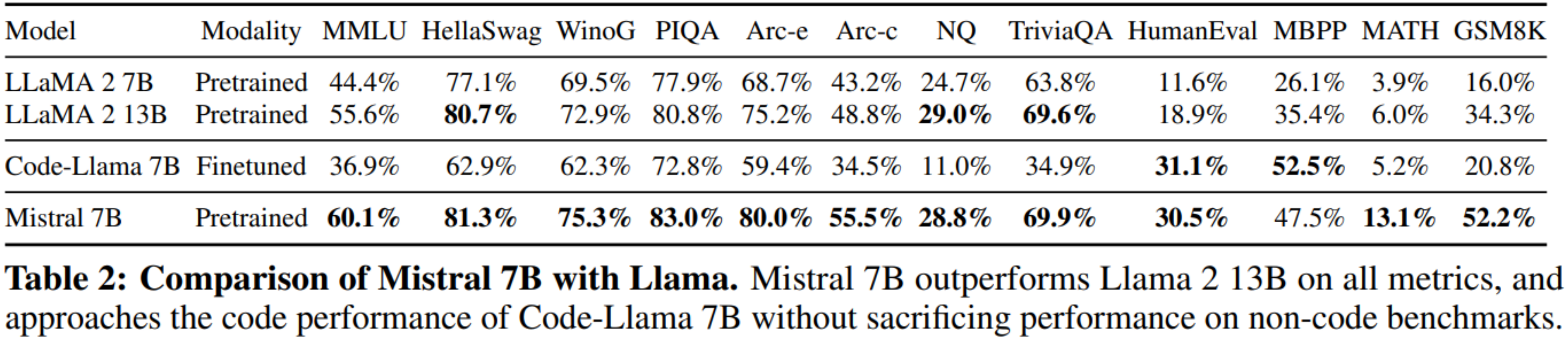
Mistral 7B : One-sided window attends to w tokens on the right side

Beltagy *et al.*, 2020 [2004.05150] Longformer: The Long-Document Transformer
Jiang *et al.*, 2023 [2310.06825] Mistral 7B

I. Mistral 7B – Results

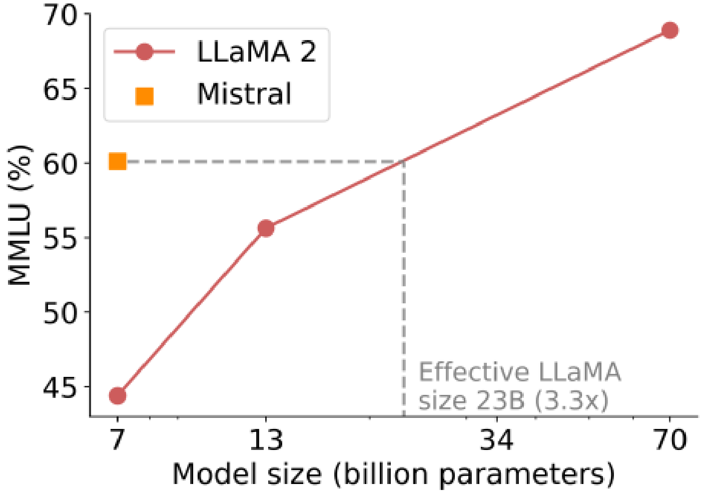
Estimated equivalent model size for LLaMA2 was 23B for MMLU **(3.3X more larger)**

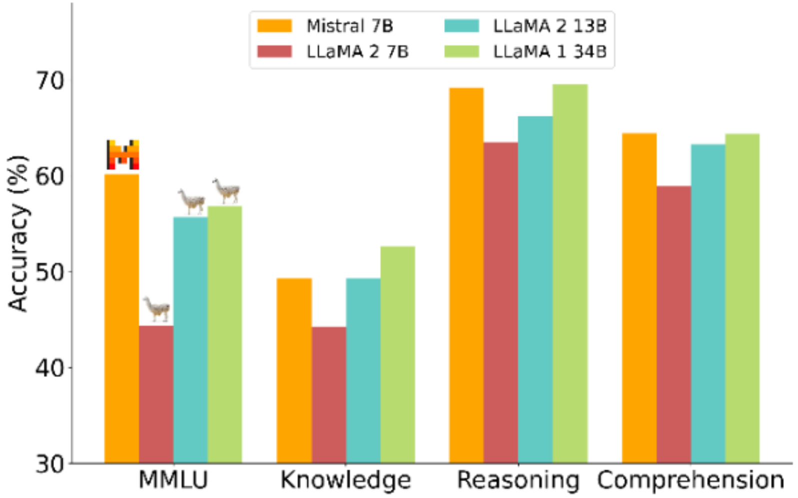
- Of the evaluated benchmarks, MMLU is the only benchmark with medical related knowledge.



Model	Modality	MMLU	HellaSwag	WinoG	PIQA	Arc-e	Arc-c	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	Pretrained	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	Pretrained	55.6%	**80.7%**	72.9%	80.8%	75.2%	48.8%	**29.0%**	**69.6%**	18.9%	35.4%	6.0%	34.3%
Code-Llama 7B	Finetuned	36.9%	62.9%	62.3%	72.8%	59.4%	34.5%	11.0%	34.9%	**31.1%**	**52.5%**	5.2%	20.8%
Mistral 7B	Pretrained	**60.1%**	**81.3%**	**75.3%**	**83.0%**	**80.0%**	**55.5%**	**28.8%**	**69.9%**	**30.5%**	47.5%	**13.1%**	**52.2%**

Table 2: Comparison of Mistral 7B with Llama. Mistral 7B outperforms Llama 2 13B on all metrics, and approaches the code performance of Code-Llama 7B without sacrificing performance on non-code benchmarks.





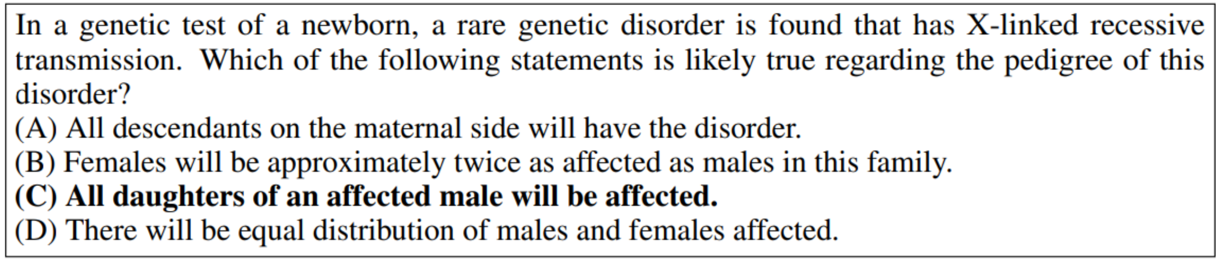
Jiang *et al.*, 2023 [2310.06825] Mistral 7B

II. Medical Language Model Benchmark

There are some publicly available evaluation dataset that are (partly) related to medicine or clinical research.

MMLU

- Popular aggregated knowledge Intensive QA
  - 57 tasks (9 tasks related to medicine)
  - College Medicine, Professional Medicine, Clinical knowledge, Anatomy, Medical Genetics, College biology (MedPaLM2 were tested with these tasks), High school biology, Virology, Nutrition
  - Measures knowledge acquired by language model using 4-way multiple-choice questions (MCQ)



In a genetic test of a newborn, a rare genetic disorder is found that has X-linked recessive transmission. Which of the following statements is likely true regarding the pedigree of this disorder?
(A) All descendants on the maternal side will have the disorder.
(B) Females will be approximately twice as affected as males in this family.
(C) All daughters of an affected male will be affected.
(D) There will be equal distribution of males and females affected.

Figure 23: A College Medicine example.

Other Benchmarks

- The following three benchmarks are mostly used
- MedQA (USMLE)
  - 5-way MCQ from US Medical License Exams
  - Focused on diagnosis
- MedMCQA
  - Mock and past exams from two Indian medical school exams
  - 4-way MCQ on 21 subjects
- PubMedQA
  - Biomedical QA dataset from PubMed abstracts
  - Answer research questions with yes/no/maybe
  - Closed domain

Benchmark	# of Subjects	Task	# of QA sets
MMLU-Medical	9	QA with 4 choices	**1,871**
MedQA	N/A	QA with 5 choices	**1,273**
MedMCQA	21	QA with 4 choices	**4,183**
PubMedQA	N/A	Abstract + QA with 3 choices	**500**

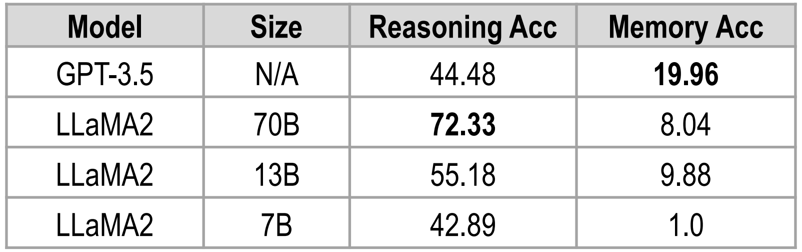
Hendrycks *et al.*, 2020 [2009.03300] Measuring Massive Multitask Language Understanding

II. Medical Language Model Benchmark

We need benchmarks beyond MCQ for medicine

Med-HALT

- Medical Domain Hallucination Test
- Reasoning Hallucination
  - False Confidence
    - Checking if the answer is correct or not
    - Model’s ability to reason the validity of the answer
  - None of the above (NOTA)
    - Correct answer is replaced with NOTA
    - Model’s ability to distinguish the irrelevant information
  - Fake Questions
    - Model’s ability to distinguish the irrelevant question
- Memory Hallucination
    - Model’s ability to recall



Model	Size	Reasoning Acc	Memory Acc
GPT-3.5	N/A	44.48	**19.96**
LLaMA2	70B	**72.33**	8.04
LLaMA2	13B	55.18	9.88
LLaMA2	7B	42.89	1.0

MedBeyMCQ

- Working on creating a novel benchmark
  - Contain novel QA sets that are manually made
- 11 Subjects (7 novel)
  - **Biomedical Engineer**, Clinical Psychologist, **Clinical Laboratory Scientist**, General Practitioner, **Occupational Therapist**, Optician, **Paramedic,** Pharmacist, **Physiotherapist**, Radiologist, **Speech-language pathologist**
- Types of tasks
  - MCQ
  - MCQ with explanation
  - Hallucination (Med-HALT approach)
  - Matching
  - Short Answer
  - Fill-in-the-blank

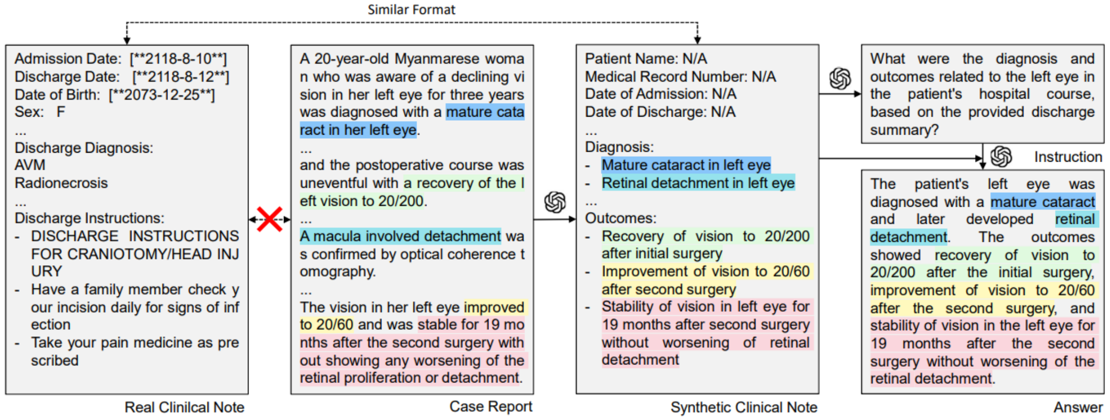
Pal *et al.*, 2023 [2307.15343] Med-HALT

III. Medical Language Models

2 Medical Related LLMs with LLaMA2 7B model as the foundation model

Asclepius

- Released on 6 September
- InstructionTuned on synthetic 158k synthetic EHR
- They also plan to release Asclepius-R which is trained on 57k real clinical notes from the MIMIC-III dataset (will be available at Physionet).



Admission Date: [**2118-8-10**]
Discharge Date: [**2118-8-12**]
Date of Birth: [**2073-12-25**]
Sex: F
...
Discharge Diagnosis:
AVM
Radionecrosis
...
Discharge Instructions:
- DISCHARGE INSTRUCTIONS FOR CRANIOTOMY/HEAD INJ URY
- Have a family member check y our incision daily for signs of inf ection
- Take your pain medicine as pre scribed

Real Clinilcal Note

A 20-year-old Myanmarese woma n who was aware of a declining vi sion in her left eye for three years was diagnosed with a mature cata ract in her left eye.
...
and the postoperative course was uneventful with a recovery of the l eft vision to 20/200.
...
A macula involved detachment wa s confirmed by optical coherence t omography.
...
The vision in her left eye improved to 20/60 and was stable for 19 mo nths after the second surgery with out showing any worsening of the retinal proliferation or detachment.

Case Report

Patient Name: N/A
Medical Record Number: N/A
Date of Admission: N/A
Date of Discharge: N/A
...
Diagnosis:
- Mature cataract in left eye
- Retinal detachment in left eye
...
Outcomes:
- Recovery of vision to 20/200 after initial surgery
- Improvement of vision to 20/60 after second surgery
- Stability of vision in left eye for 19 months after second surgery without worsening of retinal detachment

Synthetic Clinical Note

What were the diagnosis and outcomes related to the left eye in the patient's hospital course, based on the provided discharge summary?

Instruction

The patient's left eye was diagnosed with a mature cataract and later developed retinal detachment. The outcomes showed recovery of vision to 20/200 after the initial surgery, improvement of vision to 20/60 after the second surgery, and stability of vision in the left eye for 19 months after the second surgery without worsening of the retinal detachment.

Answer

Figure 2: The first column (far left) is a part of the real discharge summary taken from MIMIC-III (Johnson et al., 2016). The second column is a case report from PMC-Patients (Zhao et al., 2023), and the third is the synthetic discharge summary created from this case report. Initially, the case report did not resemble the real clinical note in

MediTron

- Released on 27 November
- Continued pretraining on PubMed papers and Medical Guidelines
- Finetuned with MedQA, MedMCQA, PubMedQA
- They also have 70B model

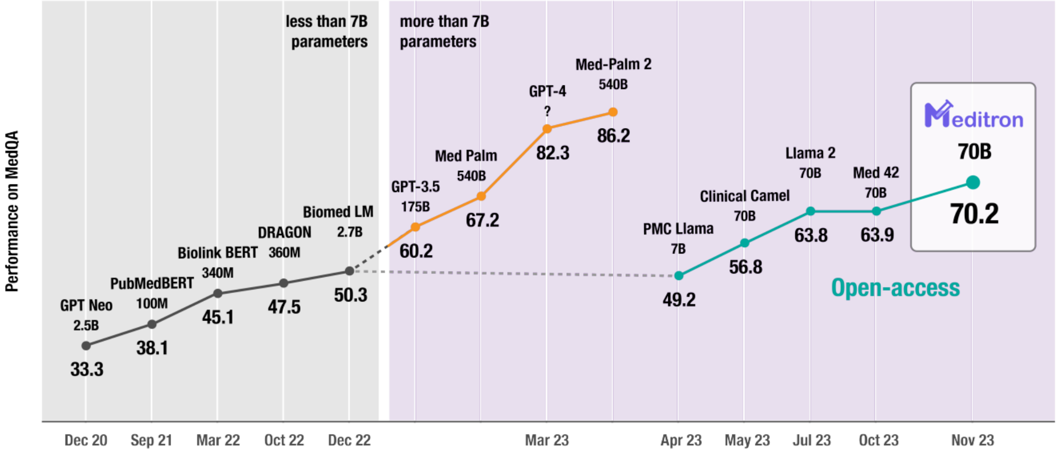


Figure 1: **MEDITRON-70B's performance on MedQA** MEDITRON-70B achieves an accuracy of 70.2 % on USMLE-style questions in the MedQA (4 options) dataset.

Kweon *et al.*, 2023 [2309.00237] Asclepius
Chen *et al.*, 2023 [2311.16079] MEDITRON-70B

III. Medical Language Models

Benchmark Results

From MediTron paper – After Finetuning

Benchmark	Mistral-7B	LLaMA2-7B	MediTron-7B
MMLU-Medical (1,862)	55.8	56.3	55.6
MedQA	32.4	44.0	47.9
MedMCQA	40.2	54.4	59.2
PubMedQA	17.8	61.8	74.4
Avg.	36.6	54.1	59.3

0-Shot Evaluation

Benchmark	Mistral-7B	LLaMA2-7B	MediTron-7B	Asclepius-LLaMA2-7B
MMLU-Medical (1,871)	67.1	40.8	35.7	39.2
MedQA	45.0	27.6	22.0	26.0
MedMCQA	49.5	36.3	31.2	33.5
PubMedQA	59.8	56.0	24.4	61.0
Avg.	55.4	40.2	28.3	39.9

MedBeyMCQ 5-Shot Evaluation

Benchmark	Mistral-7B	LLaMA2-7B	MediTron-7B	Asclepius-LLaMA2-7B
Biomedical Engineer	61.6	30.3	36.2	30.2
Clinical Psychologist	61.8	28.4	35.6	27.9
Avg.	61.7	29.4	35.9	29.1