

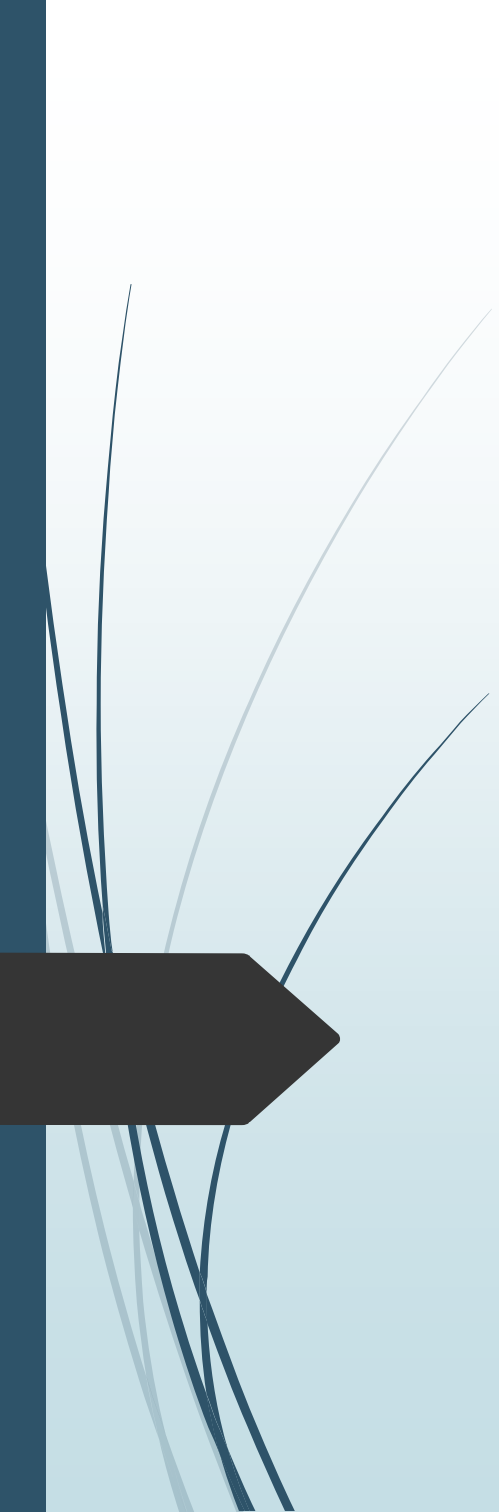
Разработка прототипа LLM-помощника по медицине и биологии



Петров
Артем



Зайчиков
Владислав



Предлагаемый нами LLM-помощник по медицине и биологии может быть интересным инструментом для получения некоторых общих сведений на указанном домене знаний.

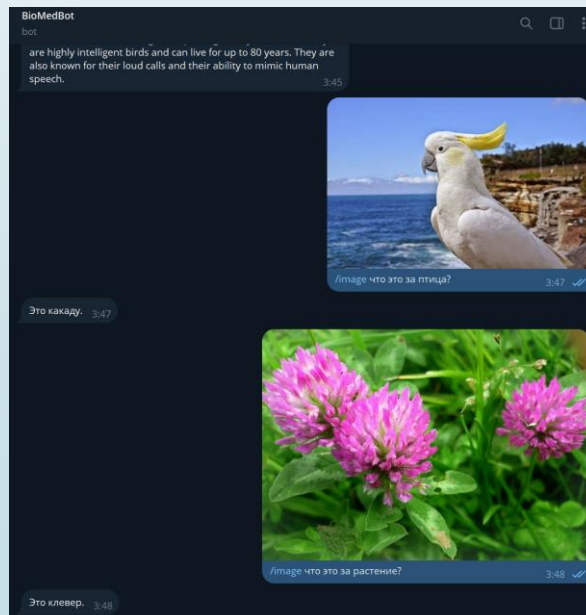
Нашей **целью** было создание своего рода рубрикатора по темам медицины и биологии, который может погрузить человека, далекого от этих областей, к дальнейшему изучению предмета. Разумеется, **рекомендации от нейросетей не могут претендовать на полноту и достоверность.**

Задачи:

- научиться работать с большими языковыми моделями
- создать прототип работоспособного приложения
- проверить качество выдаваемых ответов

В России не хватает подобных ботов, особенно более узконаправленных, а их внедрение могло бы существенно улучшить доступность и качество консультаций в области медицины и биологии.

Боты обладают огромным потенциалом: быстрые ответы, **доступные круглосуточно**, существенно упростили бы процесс консультирования и **первичной** обработки информации.



Преимущества:

- хорошие метрики качества для Gemini
- бесплатная для конечного пользователя
- свободные лимиты (120 запросов в минуту)
- не требует больших вычислительных мощностей на стороне разработчика (нас)

ChatGPT API limits:


ChatGPT API has **two rate limits: RPM (requests per minute) and TPM (tokens per minute)** ^{1 2 3}. The free trial users usually have a 20-request limit per minute, with 150,000 tokens per minute ². The rate limits for different API functionalities are as follows ³:

- Audio: 50 RPM
- Text & Embedding: 60 RPM, 250,000 TPM
- Chat: 60 RPM, 60,000 TPM
- Edit: 20 RPM, 150,000 TPM
- Image: 50 images / min

Gemini API limits

For public API entry points, we limit requests to **120 requests per minute**, and recommend that you do not exceed 1 request per second. For private API entry points, we limit requests to 600 requests per minute, and recommend that you not exceed 5 requests per second.

[Gemini REST API Reference](#)



T ▲	Model ▲	Average ▲
◆	mlabonne/Daredevil-8B-abliterated	70.86
○	ChenWeiLi/Med-ChimeraLlama-3_10k	70.8
●	Gemini-1.0	70.79
◆	shanchen/llama3-8B-slerp-biomed-chat-chinese	70.78
○	ChenWeiLi/Med-ChimeraLlama-3_1k_10_epoch	70.74
◆	shanchen/llama3-8B-slerp-med-chinese2	70.71
◆	winninghealth/WiNGPT2-Llama-3-8B-Chat	70.57
◆	HPAI-BSC/Llama3-Aloe-8B-Alpha	70.46
◆	probemedicalandyonseimailab/medllama3-v4	70.44
◆	Kukedlc/NeuralLLaMa-3-8b-DT-v0.1	70.43
■	Kukedlc/NeuralLLaMa-3-8b-ORPO-v0.3	70.4
◆	bongbongs/NewMes-v3	70.32

Наш бот Gemini-1.5, как и более продвинутая модель Gemini Pro, могут успешно справляться с задачами по пониманию и рассуждению научной биомедицинской литературы (PubMedQA) и применению клинических знаний и навыков принятия решений (MMLU Clinical Knowledge subset).

Model	Average	MedMCQA	MedQA	MMLU Anatomy	MMLU Clinical Knowledge	MMLU College Biol
ChenWeili/Med-Chimerallama-3_10k	70.8	61.15	62.14	65.93	77.36	76.39
Gemini-1.0	70.79	54.3	58	66.7	76.7	88

Источник



<https://huggingface.co/blog/leaderboard-medicalllm>

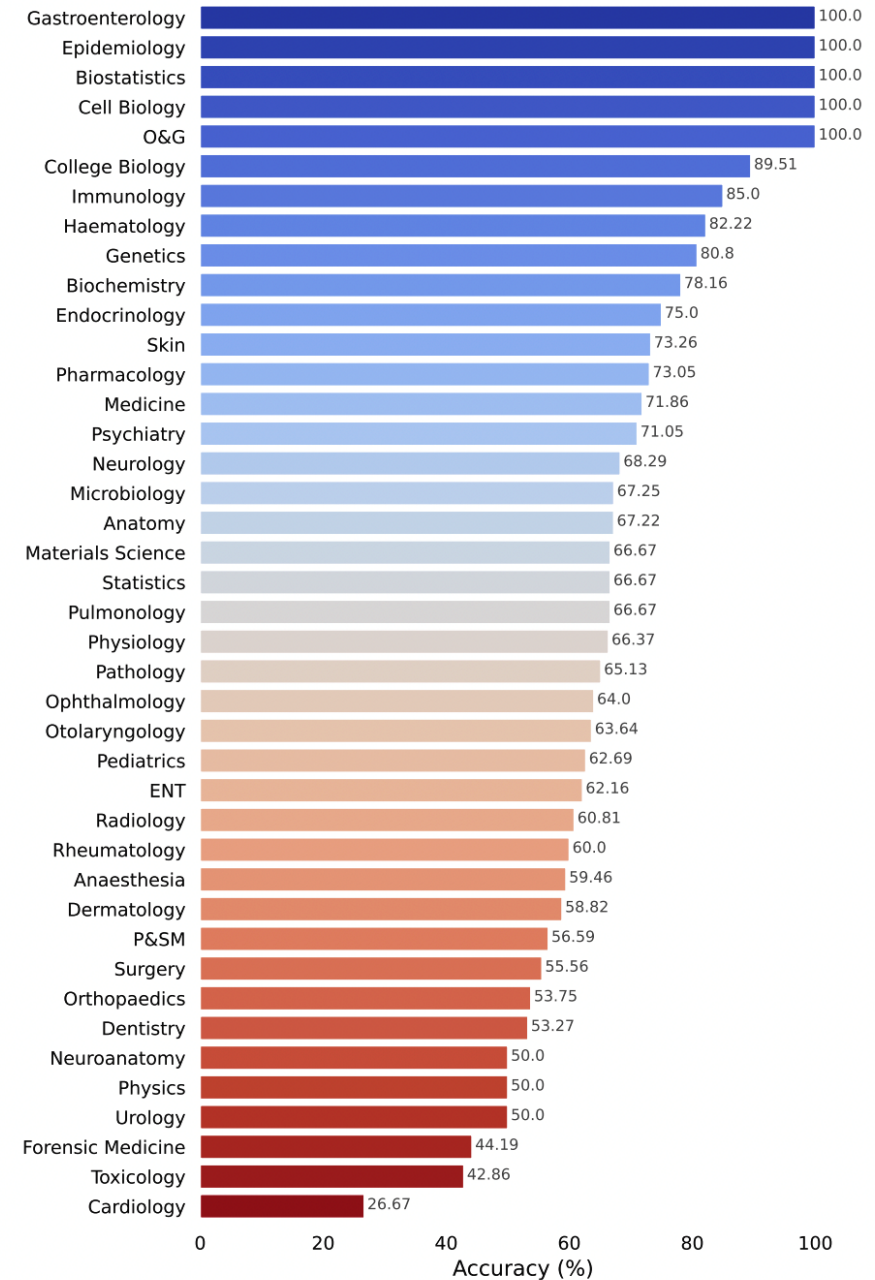
Fine-tuned Gemini для задач биомедицины



<https://research.google/blog/advancing-medical-ai-with-med-gemini/>

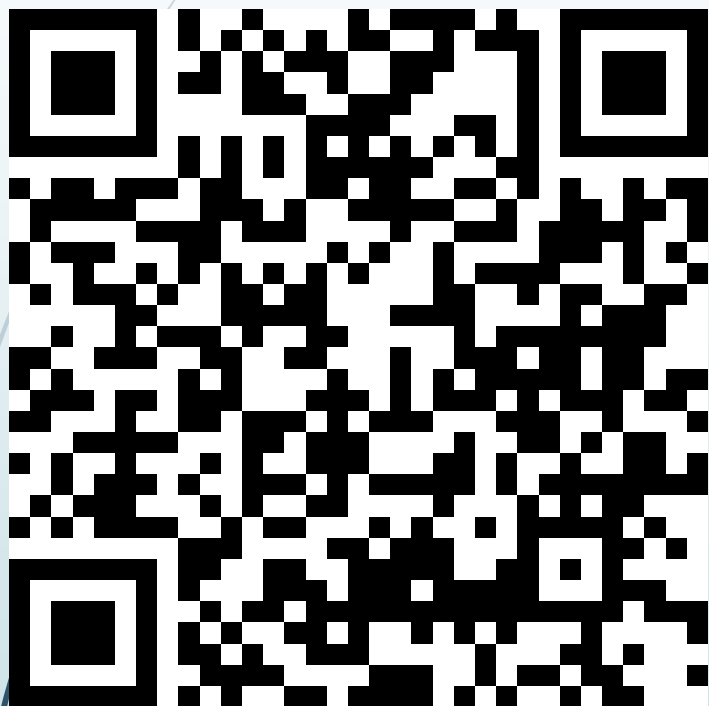
Отмечается, что Gemini Pro демонстрирует разную точность в зависимости от подраздела знаний:

- высокую — биостатистика, клеточная биология и акушерство/гинекология
- среднюю или низкую — анатомия, кардиология и дерматология



Спасибо за внимание!

Репозиторий:



<https://github.com/wlcmtunknwndth/FCSxVK/tree/dev>

Ссылка на чат с ботом:



https://t.me/FCSxVK_MedAI_bot