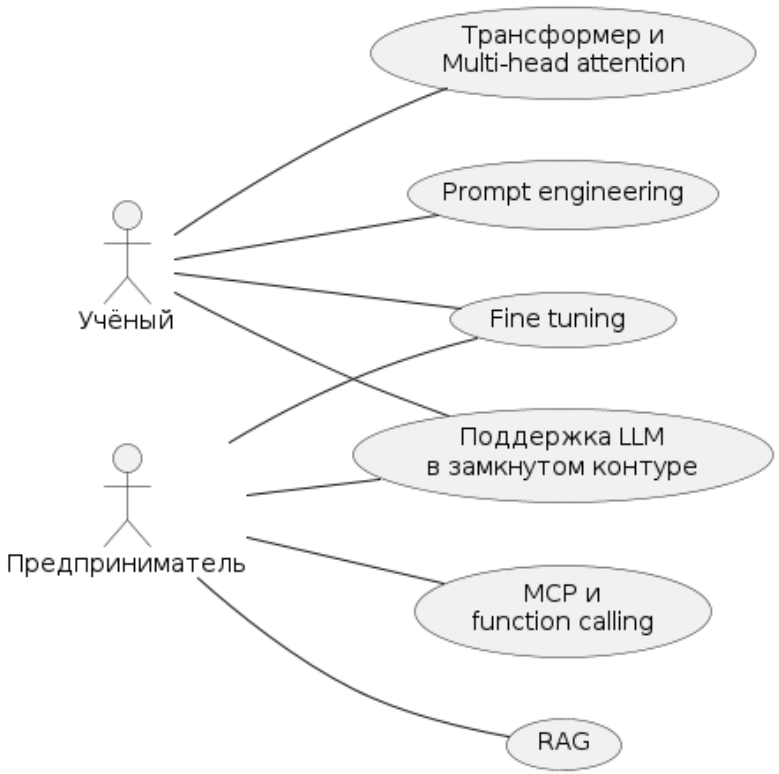
LLM между университетом и региональными компаниями

Михаил Пожидаев

21 ноября 2025 г.

Роль университета

Между учёными и предпринимателями



MCP и RAG

Комплект хороших новостей для предпринимателя

MCP — радость и грусть:

- прикрутили MCP-сервер, приложение сразу стало AI-poweered, а LLM заниматься и не пришлось;
- воспользоваться толком всё равно никто не смог, поскольку ни у кого нет Claude Desktop.

Function calling — разговор для серьёзных:

- поддержка reasoning заставляет всех беспокоиться, поэтому проектируем набор инструментов для экспорта в LLM;
- хост для OpenAI API или Anthropic API — вопрос выбора между капитальными или операционными расходами.

GPT-OSS: капелька радости в варианте 20b и одновременно капелька безысходности в варианте 120b.

Multi-head attention

Последняя надежда на публикацию

$$\alpha(Q, K, V) = \text{softmax}(\frac{Q\,K^T}{\sqrt{d}})\,V$$

Выражение внутреннего внимания зарядило интеллектуальной магией языковые модели, но его возможности всё ещё до конца неисследованы.

- Можем ли применять для обработки данных, отличных от текстовых?
- Можем ли применять для предварительного кодирования текста и решения классических лингвистических задач?

Prompt engineering

Прекрасная точка входа в исследования LLM, оставляющая немалый простор для научной новизны.

Fine tuning

И поддержка замкнутого контура

Fine tuning и языковое моделирование:

- для получения новых навыков у моделей, ориентированных на специализированные предметные области;
- для лингвистических исследований с тренировкой искусственно построенными данными.

Мы против!

Считаем большой ошибкой тренировку с целью передачи модели фактов для базы знаний — для этого есть RAG.

Утешающие технологии:

- PEFT + LoRA и Sparse mixture of experts;
- Native sparse attention и пр.

Спасибо за внимание!

Домашняя страница: https://marigostra.ru
Канал в Телеграм: https://t.me/MarigostraRu