Дообучение LLM

Assistant LLM

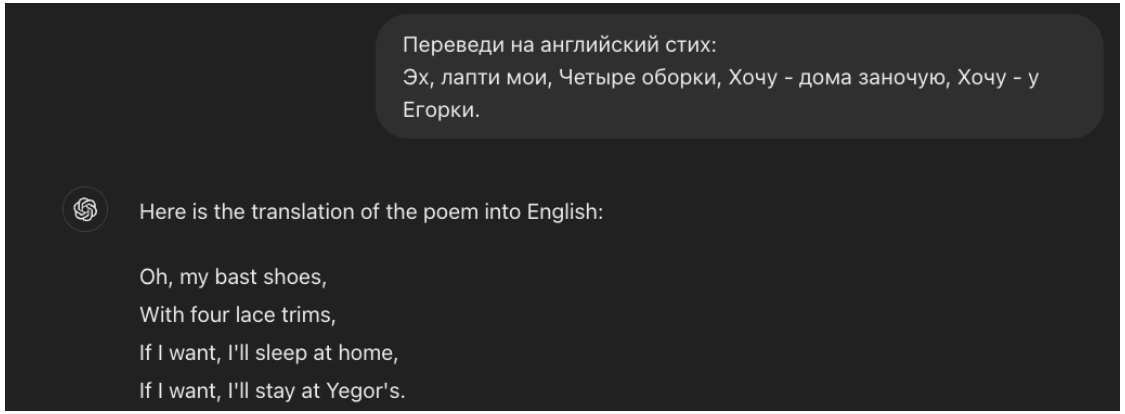
Assistant LLM (Instruction LLM) — LLM, обученная на инструкциях, на основе которых она дает ответы на любые вопросы. Таких моделей довольно много — можно подобрать себе наиболее подходящую (открытую или коммерческую), например, LLama, Mixtral, GPT4.

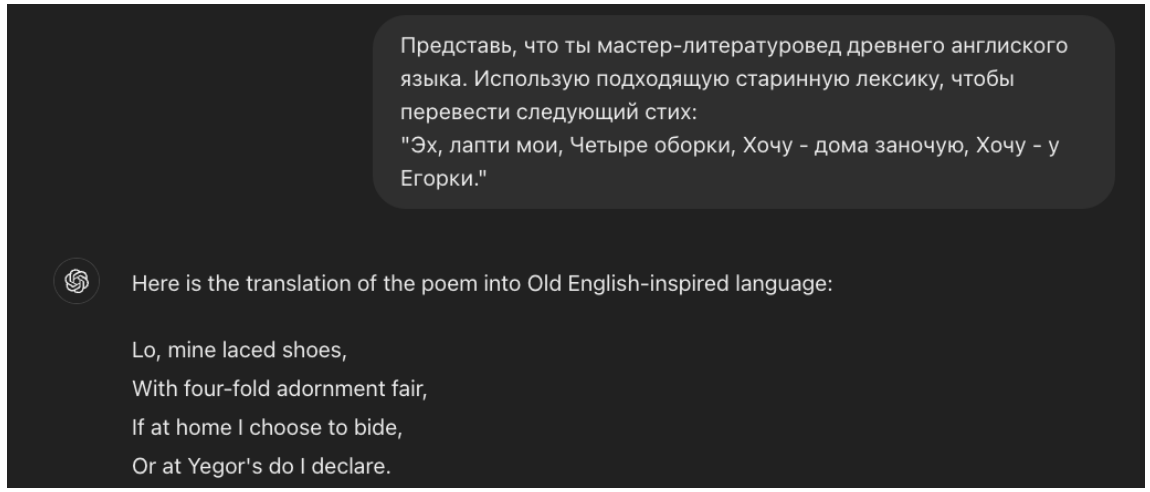
Адаптация LLM под свои нужды

- Prompt Engineering
- RAG
- Fine-tuning

Prompt Engineering

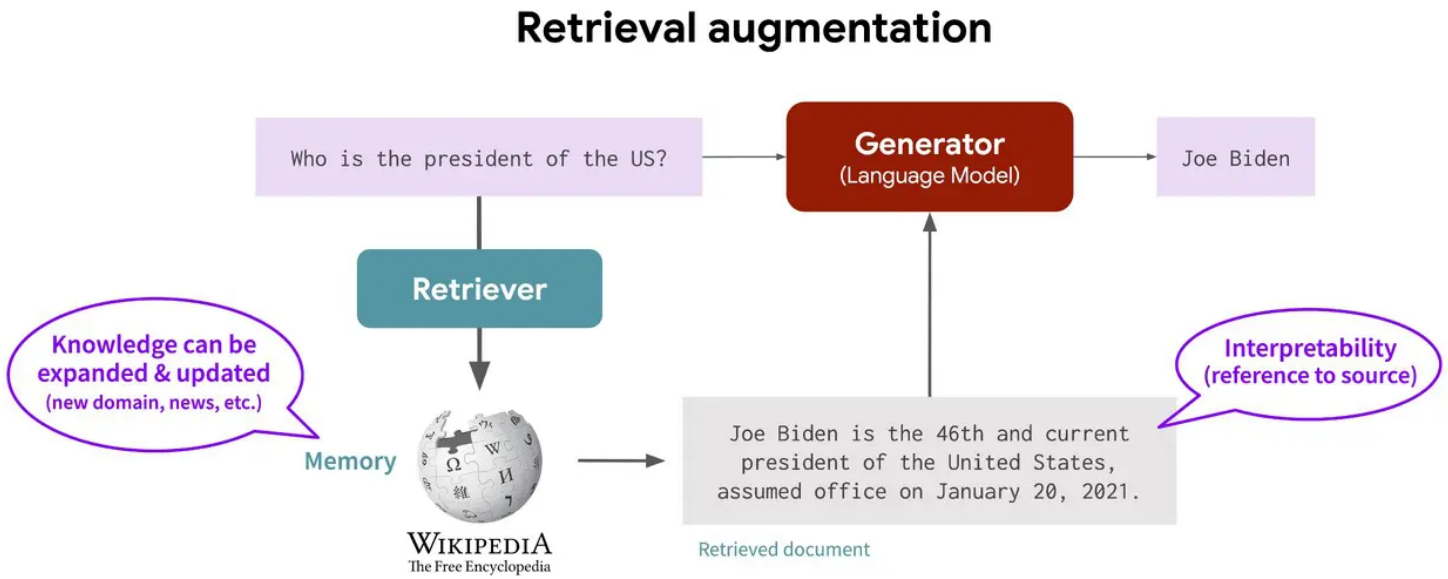
Prompting — набор техник и приемов для подачи специальных инструкций в LLM так, чтобы она отвечала определенным образом. Например, инструкция для обращения модели к клиенту по имени и отчеству или всех ее ответов с конкретной начальной фразой. От качества обучения модели и степени понимания инструкции зависит соответствие ответов заданному промпту.





RAG

RAG (Retrieval Augmented Generation) — набор техник для подачи в модель дополнительных знаний, в основном через поиск релевантных текстов и их добавление в промпт. Например, модель, обученная на данных до 2023 года, не знает о фичах айфона 2025 года. Но если мы подадим ей на вход текст с описанием нового флагмана Apple — она применит знания оттуда и не выдумает своей ответ (не будет галлюцинировать). От качества обучающего корпуса модели зависит степень ее обращения к поданному тексту.



Fine-tuning

Fine-tuning — дообучение готовой LLM на целевых данных для наиболее эффективной работы над конкретной задачей. Это первый шаг при превращении LLM в Assistant LLM.
Supervised fine-tuning (SFT) — дообучение Assistant LLM на корпусе высококачественных инструкций и ответов.

Fine-tuning

- Full fine-tuning
- Parameter Efficient fine-tuning (PEFT)

Parameter Efficient fine-tuning (PEFT)

- LoRA
- QLoRA

LoRA

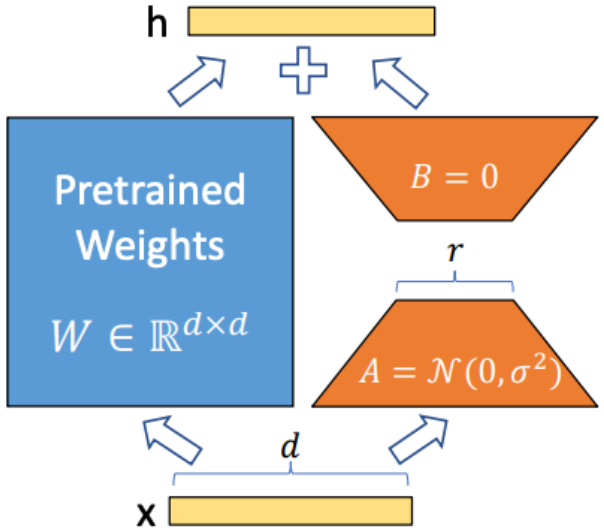


Figure 1: Our reparametrization. We only train A and B.

Линейный слой

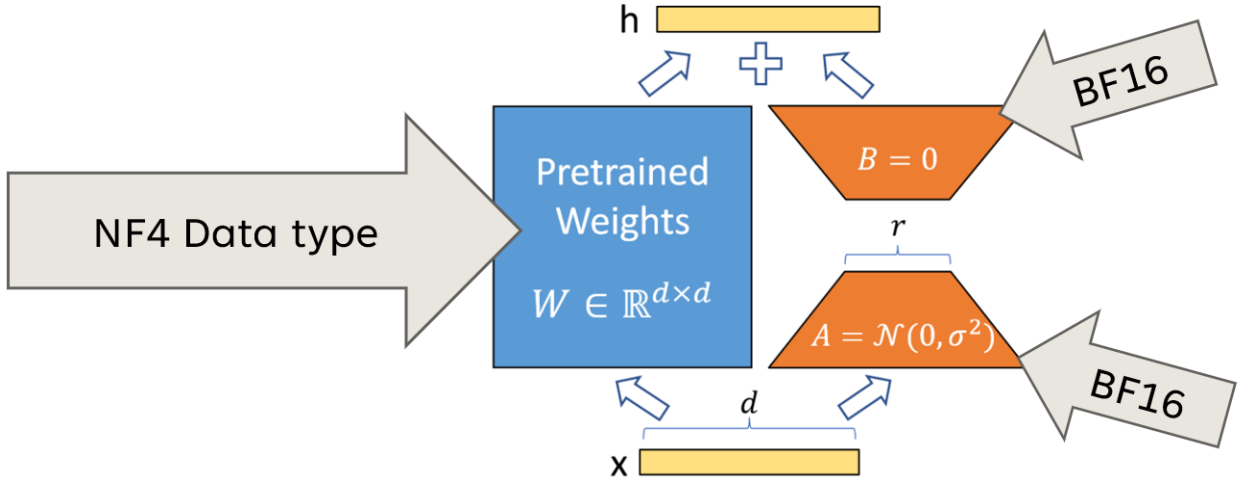
$$h = Wx$$

Линейный слой после добавления обучаемых матриц A и B

$$h = Wx + BAx$$

При этом матрица W имеет размерность [d, d], а матрицы A и B — размерности [d, r] и [r, d] соответственно, где r << d
(обычные значения d — это 8, 16, 32).
Таким образом, мы обучаем только 2 * r * d параметров вместо d * d

QLoRA


h
BF16
$B = 0$
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
NF4 Data type
r
$A = \mathcal{N}(0, \sigma^2)$
BF16
d
x

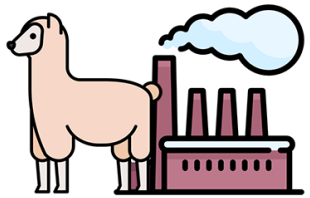
Библиотеки, модели, датасеты

Hugging Face

🤗 **Transformers**

Фреймворки для дообучения LLM

unsloth



LLaMA-Factory

Easy and Efficient LLM Fine-Tuning

Использованные материалы

- Русскоязычный пост о fine-tuning LLM
- Видео о дообучении LLM с помощью Unsloth