ti&m

Open Source LLMs: What we've learned so far

Matthias Egli, Senior Software Engineer Machine Learning

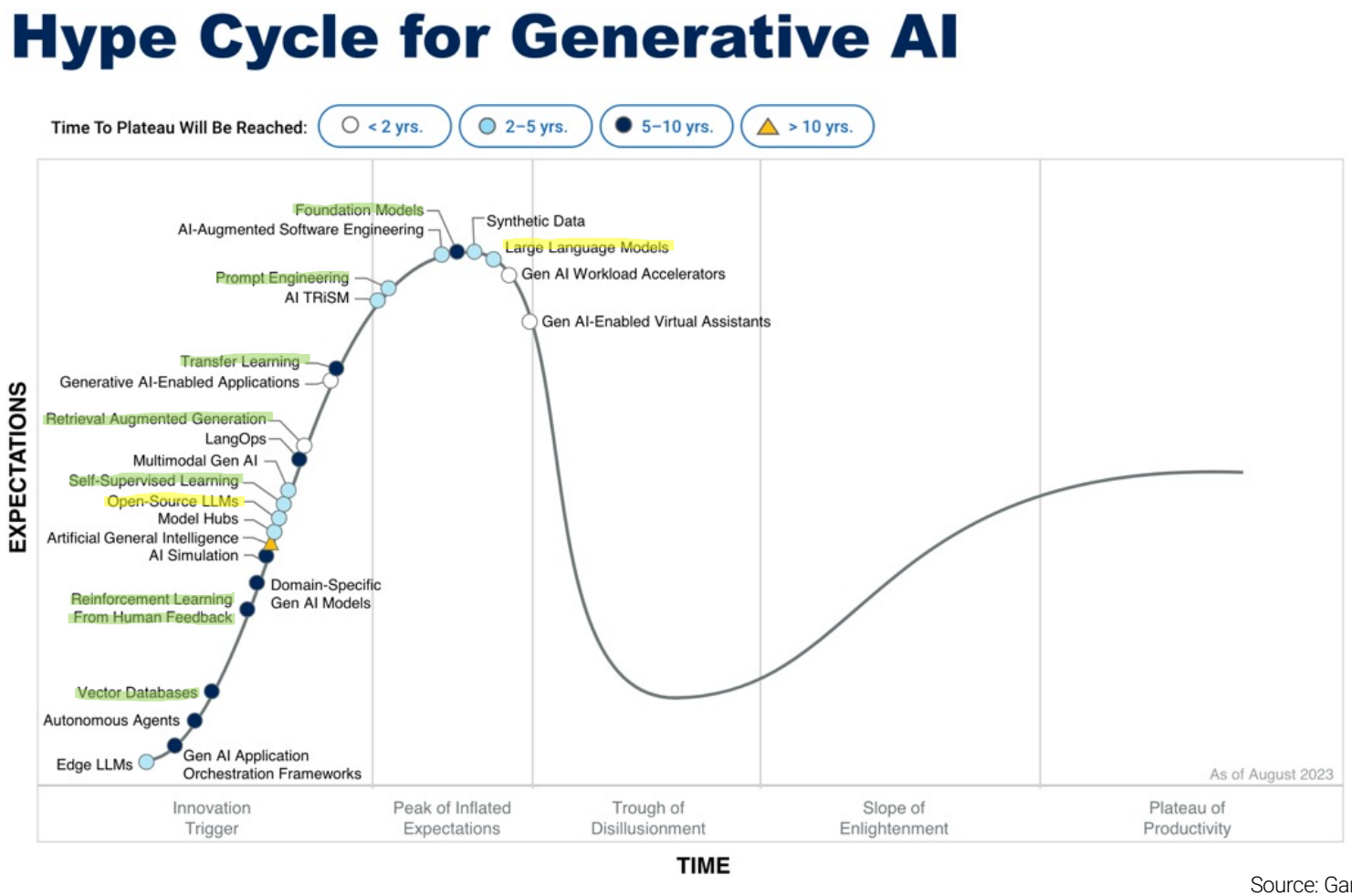
Basel, June 2024

ti&m

Who has used LLMs?

Who has used open-source LLMs?

Who has fine-tuned or augmented an open-source LLM with unseen data?



Source: Gartner Inc.

ti&m

Why using open-source LLMs in the first place?

DATA PRIVACY

CUSTOMIZATION

LATENCY

COST-EFFECTIVENESS

Multilingual chatbot for confidential data

ti&m

How to choose a suiting open-source LLM?

	Benchmark	Focus	Judge Type
1)	LMSYS Chatbot Arena Chiang et al. (2024)	Assess **human sentiment** towards model responses	Human
2)	MixEval Ni, Jinjie, et al. (2024)	**Real-world user queries**, achieving **0.96 ranking correlation** with chatbot arena	LLM
3)	IFEVAL Zhou, Jeffrey, et al. (2023)	Assess ability to **follow detailed instructions**	LLM
4)	Arena-Hard Li, Tianle, et al. (2024)	Assess **multi-turn conversation and instruction-following tasks** (0.89 correlation with human preferences)	LLM

→ *Evaluation on downstream task*

Two main strategies for task adoption of pre-trained LLMs

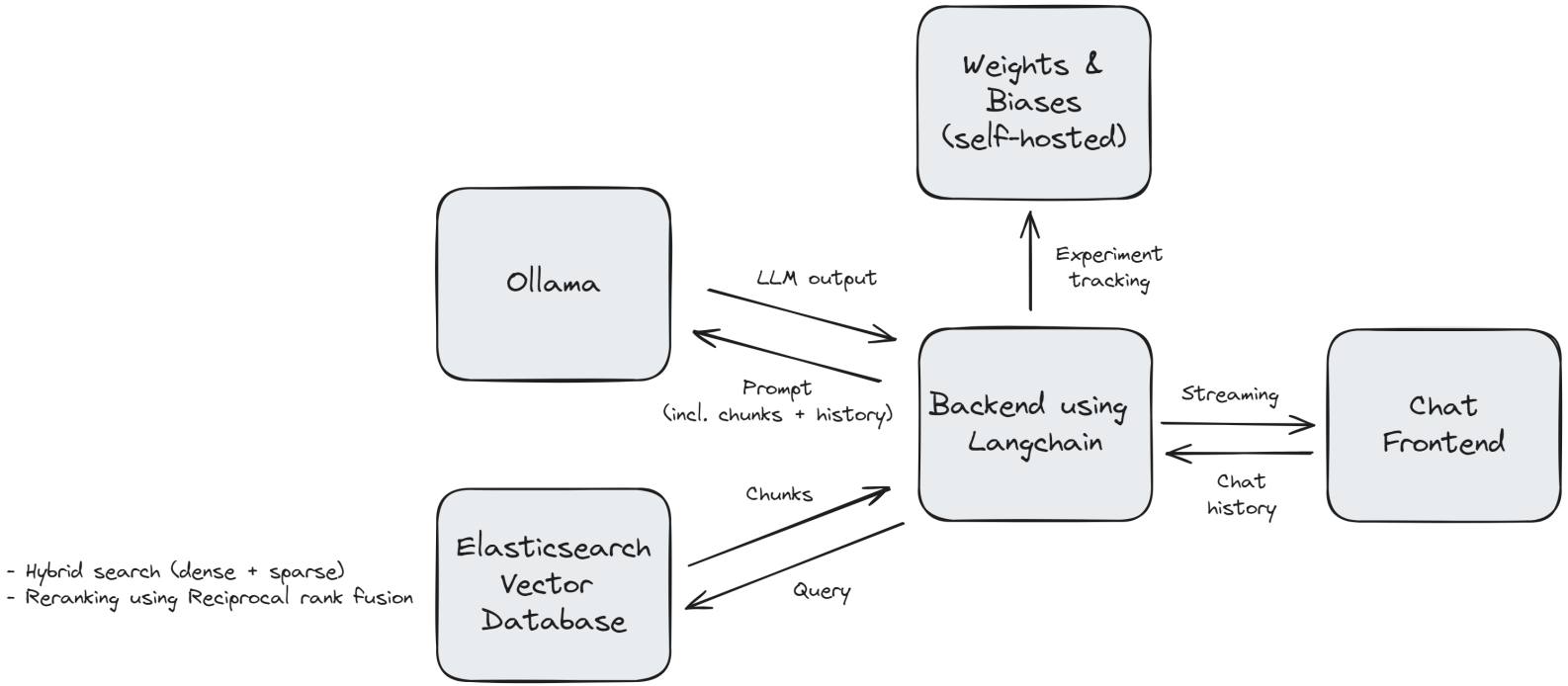
In-context learning (ICL)

+ No chat conversation data needed
+ Might alleviate hallucination
– Slower inference due to longer context
– Optimizing the retrieval part can be challenging

Supervised Fine-tuning (SFT)

+ Allows to align LLMs tone with company-specific style
– Chat conversation data needed, that is expensive to get
– More expertise required (although very standardized in the meantime)

Retrieval Augmented Generation (RAG) Architecture



ti&m

Llama-cpp vs. Ollama

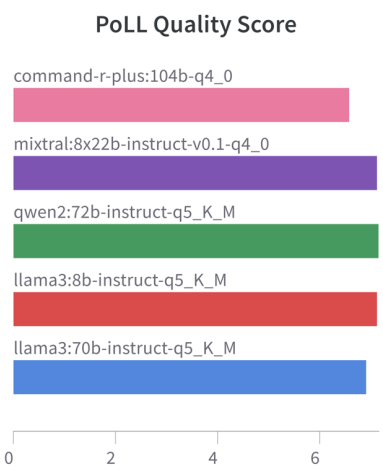
+ If you need the **features from latest releases**
– Requires dealing with prompt formats and special tokens
– Development setup for multiple developers can be tricky

+ Super easy to setup
+ **Does prompt formatting for yo**u by hiding its definition in the Modelfile
+ **Great API that streamlines development** with deployed models in **parallel**
+ Newly released models are available very quickly
+ Builds on llama.cpp

Side note: Activate flash attention to get ~1.5x throughput (if your GPU supports it)

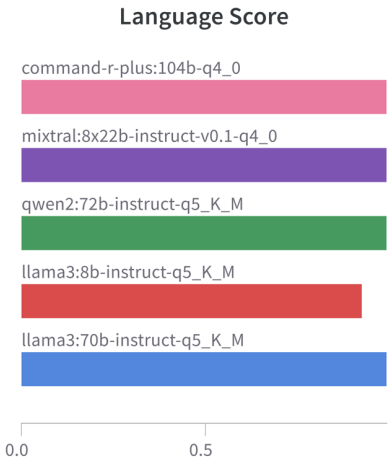
Evaluation

Metrics



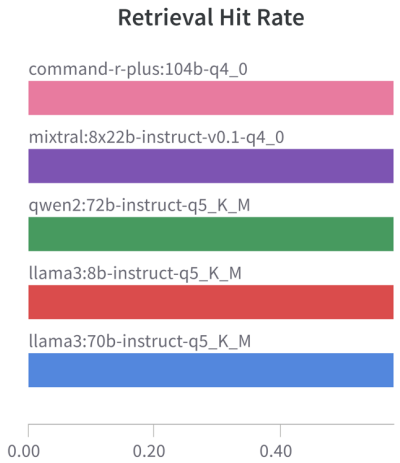
How well does the generated answer align with the ground truth answer?

- Automatic evaluation using **LLM-as-a-judge** paradigm
- **Smaller models can keep up in quality**, while offering much better latency



Does the model respond in the correct language?

- Bigger models (70B+) almost always match the language, while **smaller models occasionally (8%) miss it**



How well does the retrieval work?

- **Generative query reformulation (GQR)** has very little impact on retrieval performance

ti&m

Evaluation

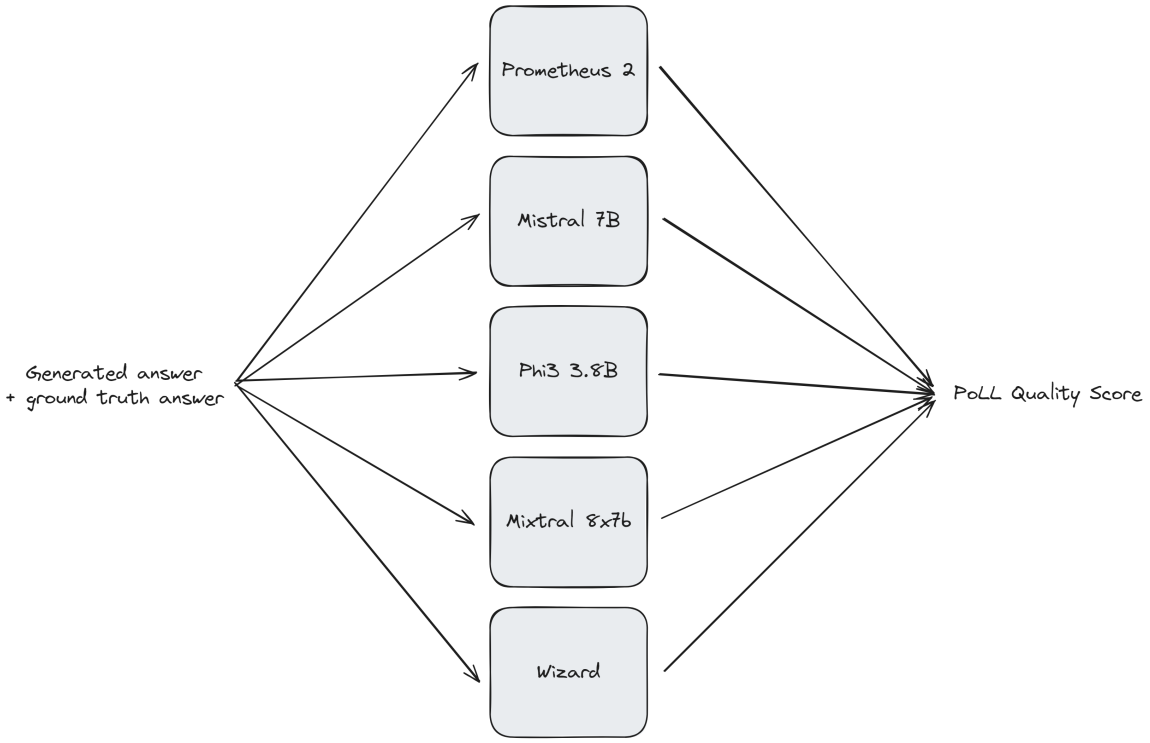
Coping with LLM-as-a-judge intra-model bias

– LLM judges have been shown to **have a positive bias towards their own answers**

– **Panel of LLM evaluators (PoLL)**: Use multiple smaller models as individual judges instead of single large model

+ Reduces intra-model bias
+ ~4 times less expensive (with 5 models)
+ Evaluate fully on premise

– May increase evaluation duration



Source: Verga, Pat, et al. (2024)

ti&m

We digitalize your company.

600+

talents

Consultants, analysts, designers, system and software engineers.

No. 1

in Switzerland

for digitalization, security and innovation projects and products.

100 %

vertical integration

of the entire IT value chain.

6

offices

Zurich, Bern, Basel, Frankfurt am Main, Dusseldorf and Singapore.

Growth

Sustainable and organic growth

Our expertise in digital banking & finance, insurance, eGovernment and public sector, transport and logistics, life sciences and industry, retail and national security makes us the right partner for you.

Top 10

of the largest owner-managed Swiss IT companies

Founded in 2005 and 100 % owner-managed since then.