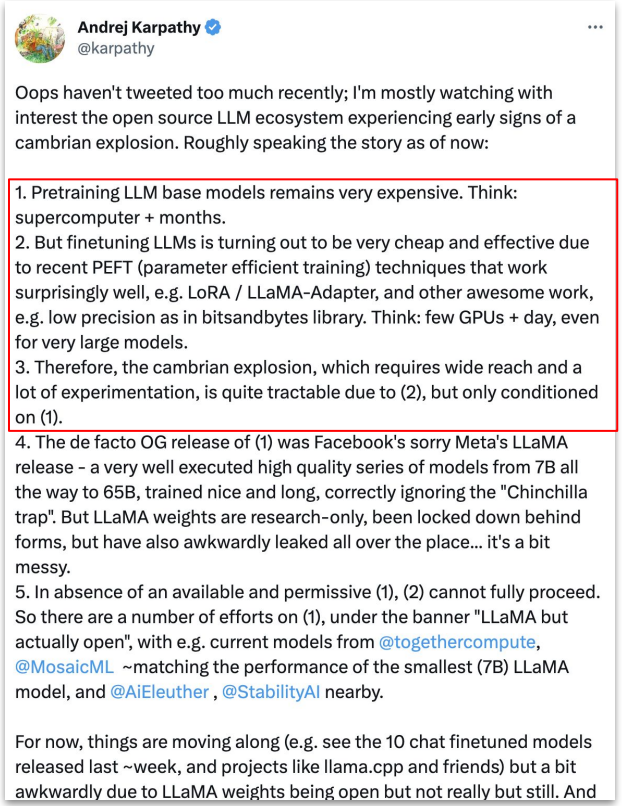
- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Nachhaltigkeit und Trainingskosten

Johannes Hötter
Christian Warmuth

Wie viel Aufwand steckt dahinter, ein LLM zu bauen?



Andrej Karpathy
@karpathy

Oops haven't tweeted too much recently; I'm mostly watching with interest the open source LLM ecosystem experiencing early signs of a cambrian explosion. Roughly speaking the story as of now:

1. Pretraining LLM base models remains very expensive. Think: supercomputer + months.
2. But finetuning LLMs is turning out to be very cheap and effective due to recent PEFT (parameter efficient training) techniques that work surprisingly well, e.g. LoRA / LLaMA-Adapter, and other awesome work, e.g. low precision as in bitsandbytes library. Think: few GPUs + day, even for very large models.
3. Therefore, the cambrian explosion, which requires wide reach and a lot of experimentation, is quite tractable due to (2), but only conditioned on (1).
4. The de facto OG release of (1) was Facebook's sorry Meta's LLaMA release - a very well executed high quality series of models from 7B all the way to 65B, trained nice and long, correctly ignoring the "Chinchilla trap". But LLaMA weights are research-only, been locked down behind forms, but have also awkwardly leaked all over the place... it's a bit messy.
5. In absence of an available and permissive (1), (2) cannot fully proceed. So there are a number of efforts on (1), under the banner "LLaMA but actually open", with e.g. current models from @togethercompute, @MosaicML ~matching the performance of the smallest (7B) LLaMA model, and @AiEleuther , @StabilityAI nearby.

For now, things are moving along (e.g. see the 10 chat finetuned models released last ~week, and projects like llama.cpp and friends) but a bit awkwardly due to LLaMA weights being open but not really but still. And

Grundtraining von GPT-3 für OpenAI: mehrere Monate auf einem Supercomputer

Finetuning von Open-Source LLMs z.B. für Versicherungsfirma: wenige Tage auf mehreren GPUs

Nachhaltigkeit und Trainingskosten

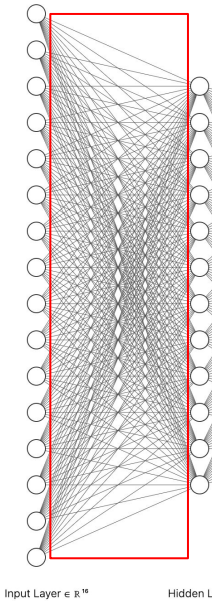
Johannes Hötter,
Christian Warmuth

openHPI

Quelle: https://twitter.com/karpathy/status/1654892810590650376?lang=en

Recap
Quantisierung von Modellen zur Ressourcenschonung

- Quantisiertes Modell, welches sogar auf einem MacBook M1 laufen kann!
- Komprimiert durch Reduzierung des Datentyps der Modellparameter.



Sinngemäß (vereinfachte Darstellung):

- Parameter im Originalzustand mit 32 Bytes:
  3,14159265
- Mit 16 Bytes wäre abbildbar:
  3,14
- Somit kann das Modell verkleinert werden, ohne zu große Performance-Einbußen in Kauf nehmen zu müssen
- Ein Modell mit 1.000.000.000 Parametern braucht grob 4GB RAM, in quantisierter Form nur 1GB.

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Nachhaltigkeit und Trainingskosten

Johannes Hötter,
Christian Warmuth

openHPI

Wie viel Aufwand steckt dahinter, ein LLM zu bauen?

Home > News > Software & Entwicklung > ChatGPT und Kenia: Schlecht bezahlte Arbeiter müssen strafbare Inhalte aussortieren

News

ChatGPT und Kenia: Schlecht bezahlte Arbeiter müssen strafbare Inhalte aussortieren

Um die Datenmengen für die Verarbeitung durch eine KI vorzubereiten, greift die Branche auf schlecht bezahlte Arbeitskräfte zurück. Sie müssen die Inhalte sichten, die keiner von uns sehen möchte.

Von **Kim Rixecker**

29.01.2023, 08:30 Uhr • 3 Min. Lesezeit

Quelle: https://t3n.de/news/chatgpt-openai-inhalte-filtern-schlecht-bezahlte-arbeiter-1528694/, aufgerufen 04. Juni 18:00

Digital Engineering · Universität Potsdam

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

Nachhaltigkeit und Trainingskosten

Johannes Hötter,
Christian Warmuth

openHPI

Recap
Ein simples Beispiel (Labeling)

Karteikarte **Vorderseite**

Karteikarte **Rückseite**

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum HPI
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Nachhaltigkeit und Trainingskosten

Johannes Hötter,
Christian Warmuth

openHPI

Recap
Ziel von LLMs: Weltwissen abbilden

Etwa:

- Wenn die Sonne scheint
- Und es gleichzeitig regnet
- Dann gibt es einen Regenbogen

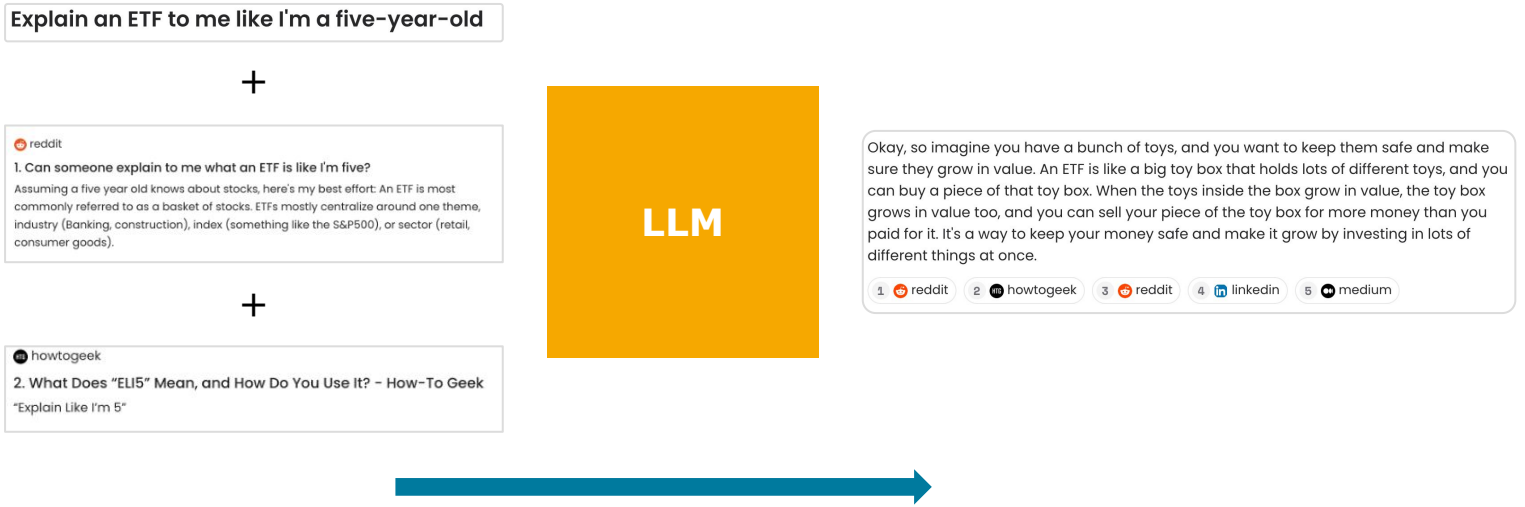
Je besser dieses Weltwissen in quantisierten Modellen (komprimiert) abgespeichert werden kann, desto mehr Ressourcen können beim Training und zur Laufzeit gespart werden.

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum HPI
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Recap
Domänenwissen über Finetuning oder In-Context Learning oder Prompt Engineering

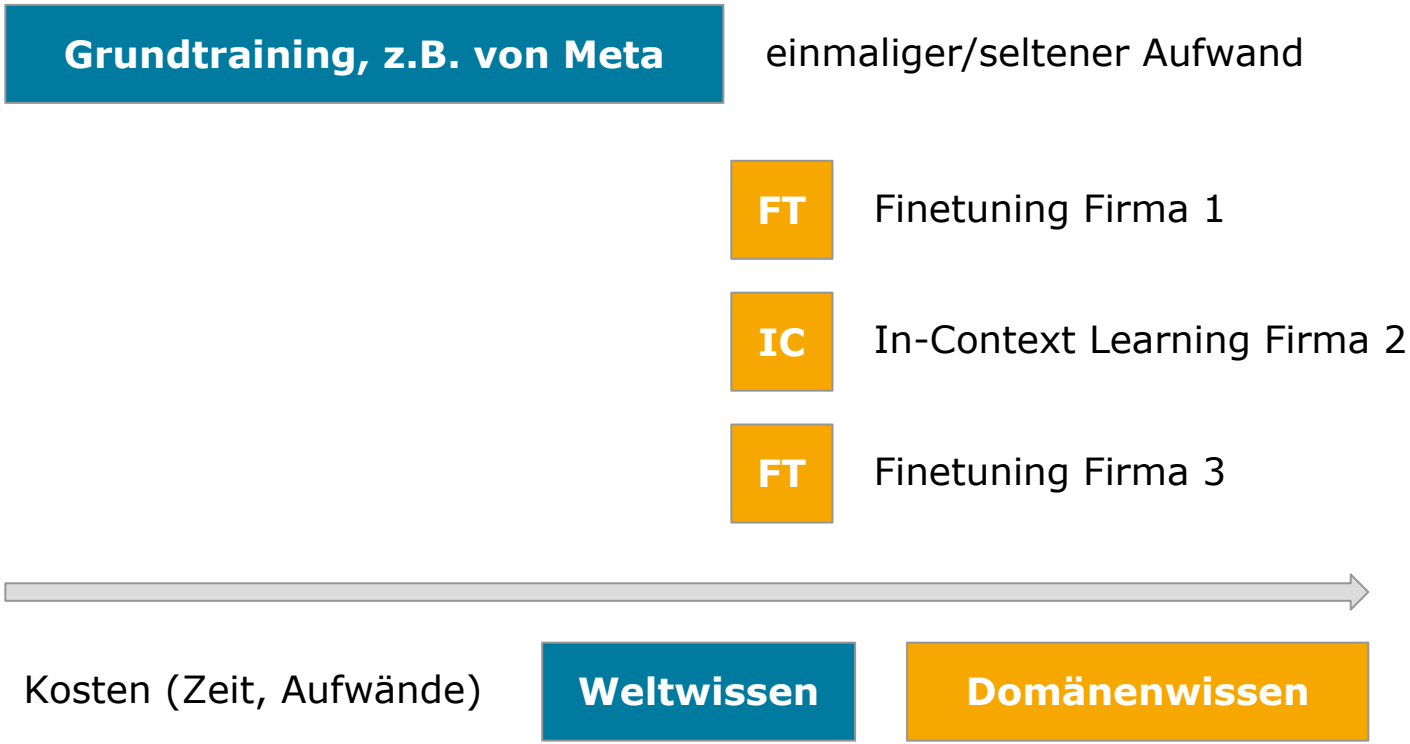


Nachhaltigkeit und Trainingskosten

Johannes Hötter,
Christian Warmuth

openHPI

Open-Source, quantisierte LLMs werden zur Nachhaltigkeit und Kosteneffizienz beitragen



HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

KI Service Zentrum HPI
by Hasso-Plattner-Institut

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Nachhaltigkeit und Trainingskosten

Johannes Hötter,
Christian Warmuth

openHPI

- openHPI: ChatGPT: Was bedeutet generative KI für unsere Gesellschaft? -

Nachhaltigkeit und Trainingskosten

Johannes Hötter
Christian Warmuth