



Wie
konkurrenzfähig
sind OpenLLMs?

ՀԻՆՉ ՕԲԵՍԴՐՄԵՆ
ԿՈՆԿՈՒՐԵՆՑԱԿԱՆԱԿՆԵՐՄԵՆ

ՄԻԾ

Semesterprojekt: OpenLLM-Evaluation

- Teil 1:
 - Aufgabenspezifikation und Anlegen eines GitHub-Verzeichnisses (bis 3. Juni 12:00 Uhr)
 - Kurzvorstellung 6. Juni im Kurs
- Teil 2:
 - Experiment-Durchführung, Analyse und Dokumentation (bis 24. Juni 12:00)
 - Ca. 15-minütige Vorstellung in einer der letzten 4 Sitzungen des Semesters
- Hinweis: Ihr Experiment muss sich von dem, das Sie im Hauptseminar durchführen, unterscheiden!

Aufgaben-, Modell- & Datenauswahl

- **Möglichst klare Definition der Aufgabe:** Auf welche Funktionalität möchten Sie das LLM testen?
- **Auswahl eines geeigneten OpenLLMs:** Weshalb haben Sie sich für das gewählte Modell entschieden?
- **Auswahl eines geeigneten Datensatzes:** Wählen Sie einen Datensatz, der repräsentativ für die Aufgabe ist und ausreichende Komplexität bietet, um die Fähigkeiten des Modells zu testen / zu trainieren.

Formulierung und Optimierung der Prompts

- Konsultieren Sie die Dokumentation der Modelle: Gibt es Strukturen, auf die die Modelle optimiert wurden (z.B. system / user / assistant).
- Wenn Sie deutsche Prompts verwenden, testen Sie auf jeden Fall auch englische.
- Testen Sie Strategien wie positive-thinking, in-context-learning (Lui et al. 2023), optimalerweise auch automatisierte Optimierung (bspw. mit DSPy, Battle et al. 2024)
- Gehen Sie iterativ vor!

Analyse und Vergleich der Ergebnisse

- Wie verhalten sich die Ergebnisse gegenüber ihren anfänglichen Erwartungen?
- Wie verhalten sich die Ergebnisse gegenüber der Performance bekannter, nicht-Open-LLMs?
- Lassen sich Muster / Trends identifizieren, was gut und was weniger gut funktioniert bzw. sich optimieren lässt?

Dokumentation im Github-Verzeichnis

Dokumentieren Sie in einer möglichst übersichtlichen Struktur:

- Aufgabenstellung, LLM-Auswahl
- Verwendete Daten
- Prompts, inkl. deren Veränderung über die Projektlaufzeit (haben Sie irgendwo nachgebessert / verändert o.ä. - dokumentieren Sie auch, aus welchem Grund bzw. auf Basis welcher Strategie Sie die Prompts entworfen haben)
- Ergebnisse (Antworten des/der LLM)
- Code, den Sie zum scrapen / prä-prozessieren / prompt-generieren / analysieren der Daten geschrieben / verwendet haben
- Ergebnisauswertung, gerne mit Grafiken