

Code Llama, a state-of-the-art large language model for coding

Danh Thien Luu (79663)

Inhaltsverzeichnis

| | | |
|---|--------------------------------|---|
| 1 | Einführung | 1 |
| 2 | Zahlen und Fakten | 2 |
| 3 | Einrichten von Code Llama | 4 |
| 4 | Durchführung | 5 |
| 5 | Vergleich mit anderen Chatbots | 6 |
| 6 | Ergebnisse | 7 |
| 7 | Fazit | 8 |
| | Referenzen | 9 |

1 Einführung

Code Llama ist ein neues State of the Art Sprachmodell, spezialisiert zur Generierung von Code und natürlicher Sprache über Code. Dazu akzeptiert sie sowohl Prompts, die Code enthalten als auch welche, die natürliche Sprache enthalten.

Entwickelt wurde sie von Meta, der Muttergesellschaft von Facebook und ist frei zu Forschungs- und Kommerziellen Zwecken nutzbar. Veröffentlicht wurde Code Llama am 24. August und ist, ein auf Programmiercode spezialisierte Version von Llama 2, welches durch ein erweitertes, längeres Training des bestehenden Code-Datensatzes entstand. [1]

Als Ergebnis, kann Code Llama Code generieren, und unterstützt zudem

In dieser Projektarbeit wird Code Llama auf die beworbenen Fähigkeiten getestet und auf die Nutzbarkeit in der echten Welt geprüft.

2 Zahlen und Fakten

Code Llama gibt es in verschiedenen Varianten:

- Default
- Instruct
- Python
- Unnatural

Diese unterscheiden sich in erster Linie durch den Datensatz, die zum Trainieren der Modelle verwendet wurde.

Die Default-Variante ist die Standardvariante mit dem Datensatz, die vorher beschrieben wurde, basiert.

Bei der Python-Variante wurde mehr Python-Code zum feintunen verwendet, somit kann dieses Modell in der Theorie besser mit Python umgehen.

Bei der Instruct Variante handelt es sich um ein Modell, welches mit menschlichen Instruktions-Datensätzen fein getunt worden ist. Man nennt dies dann aligned, dies bedeutet die Ausgabe des Modells ist konsistent zu dem, wie es ein Mensch erwarten würde, dies ermöglicht es dem Modell, z.B. auf Fragen zu antworten oder andere menschenähnliche Interaktionen zu erzeugen. Somit ist das die nutzerfreundlichste Version von Code Llama.

Es gibt auch noch eine Variante, die als Unnatural Code Llama bezeichnet wird. Diese Version wird der Öffentlichkeit leider (noch) nicht zur Verfügung gestellt. Sie schneidet im Vergleich zu den anderen Sprachmodellen, die im Research Paper verglichen wurden in allen bis auf einer Rubrik am besten ab und dürfte somit Metas mächtigstes Sprachmodell für Code sein. Erschaffen wurde es, indem Code Llama - Python anhand von 15.000 unnatürlichen Instruktionen feingetunt worden ist, also ein Datensatz, der vollkommen synthetisch und automatisiert mithilfe von anderen Sprachmodellen erzeugt wurde.

jedoch gibt es, bis auf eine kurze Erwähnung in der Vergleichstabelle, dazu keine näheren Informationen über dieses Modell.

Diese verschiedenen Variationen gibt es dann nochmal in der 7b, 13b und der 34b Version, welche die Größe des Sprachmodells beschreibt. A

Es gibt viele Sprachmodelle, oder genauer bezeichnet „Large Language Models“ (LLM), die in Konkurrenz mit Code Llama stehen wie z.B. GPT-4, die im selben bzw. anderen Bereichen besser performen können.

Dazu gibt es verschiedene Kennzahlen, um diese Performance zu evaluieren.

In der Tabelle, die im Research Paper auftaucht, kann man erkennen, dass

| Model | Size | HumanEval | | | MBPP | | |
|-----------------------|-------|--------------|--------------|--------------|--------------|--------------|--------------|
| | | pass@1 | pass@10 | pass@100 | pass@1 | pass@10 | pass@100 |
| code-cushman-001 | 12B | 33.5% | - | - | 45.9% | - | - |
| GPT-3.5 (ChatGPT) | - | 48.1% | - | - | 52.2% | - | - |
| GPT-4 | - | 67.0% | - | - | - | - | - |
| PaLM | 540B | 26.2% | - | - | 36.8% | - | - |
| PaLM-Coder | 540B | 35.9% | - | 88.4% | 47.0% | - | - |
| PaLM 2-S | - | 37.6% | - | 88.4% | 50.0% | - | - |
| StarCoder Base | 15.5B | 30.4% | - | - | 49.0% | - | - |
| StarCoder Python | 15.5B | 33.6% | - | - | 52.7% | - | - |
| StarCoder Prompted | 15.5B | 40.8% | - | - | 49.5% | - | - |
| LLAMA 2 | 7B | 12.2% | 25.2% | 44.4% | 20.8% | 41.8% | 65.5% |
| | 13B | 20.1% | 34.8% | 61.2% | 27.6% | 48.1% | 69.5% |
| | 34B | 22.6% | 47.0% | 79.5% | 33.8% | 56.9% | 77.6% |
| | 70B | 30.5% | 59.4% | 87.0% | 45.4% | 66.2% | 83.1% |
| CODE LLAMA | 7B | 33.5% | 59.6% | 85.9% | 41.4% | 66.7% | 82.5% |
| | 13B | 36.0% | 69.4% | 89.8% | 47.0% | 71.7% | 87.1% |
| | 34B | 48.8% | 76.8% | 93.0% | 55.0% | 76.2% | 86.6% |
| CODE LLAMA - INSTRUCT | 7B | 34.8% | 64.3% | 88.1% | 44.4% | 65.4% | 76.8% |
| | 13B | 42.7% | 71.6% | 91.6% | 49.4% | 71.2% | 84.1% |
| | 34B | 41.5% | 77.2% | 93.5% | 57.0% | 74.6% | 85.4% |
| UNNATURAL CODE LLAMA | 34B | 62.2% | 85.2% | 95.4% | 61.2% | 76.6% | 86.7% |
| CODE LLAMA - PYTHON | 7B | 38.4% | 70.3% | 90.6% | 47.6% | 70.3% | 84.8% |
| | 13B | 43.3% | 77.4% | 94.1% | 49.0% | 74.0% | 87.6% |
| | 34B | 53.7% | 82.8% | 94.7% | 56.2% | 76.4% | 88.2% |

Abbildung 1: Code LLama im Vergleich zu anderen LLMs [2]

3 Einrichten von Code Llama

4 Durchführung

Um die Fähigkeiten der Code- und Sprachgenerierung von Code Llama auszuprobieren und zu testen, könnten folgende Nutzzwecke Interessant sein:

- Schreiben von Code anhand einer Beschreibung in natürlicher Sprache
- Code Reviews
- Dokumentation
- Verbessern von bestehendem Code (Refactoring)
- Zusammenfassung von Code
- Potenzielle Bugs erkennen und Hilfestellung bei Debugging leisten

5 Vergleich mit anderen Chatbots

6 Ergebnisse

7 Fazit

Referenzen

- [1] Introducing Code Llama, a State-of-the-Art Large Language Model for Coding. Abgerufen 16. Oktober 2023 von <https://ai.meta.com/blog/code-llama-large-language-model-coding/>
- [2] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, und Gabriel Synnaeve. Code Llama: Open Foundation Models for Code.