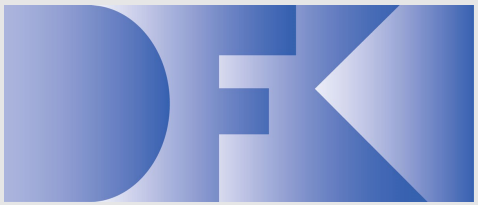
Evaluation with the Language Model Evaluation Harness for German

Konstantin Schulz[1]

[1]German Research Center for Artificial Intelligence

May 17, 2022

openGPT-X



Overview

Language Model Evaluation Harness

- introduced by EleutherAI
- for autoregressive models (GPT-like)
- only for English
- collection of 200+ datasets
- conversion of tasks
- compatible with Huggingface

Checklist

- ☑ support XGLM architecture
- ☑ add GermanQuAD (Extractive Question Answering) as closed-book QA task
- ☑ evaluate various smaller German GPT models
- ☐ reformulate GermanQuAD as few-shot abstractive QA
- ☐ add other German tasks, e.g. summarization
- ☐ support Megatron-DeepSpeed
- ☐ evaluate larger models

Sample Output of German GPT2 1.5B: Correct

Question	Prediction	Ground Truth
Welchen Preis gewann Robert Koch für die Entdeckung des Erregers von Tuberkulose?	den Nobelpreis für Medizin	Nobelpreis für Physiologie oder Medizin
Welcher König erwarb den Buckingham Palace 1761?	der englische König Georg III.	König Georg III.

Sample Output of German GPT2 1.5B: Nice Try

Question	Prediction	Ground Truth
Weswegen war die New York Stock Exchange 1888 geschlossen?	Die Börse wurde von der Regierung geschlossen, weil sie zu viele Verluste gemacht hatte.	Großer Schneesturm
In welchem weiteren Museums Haus hat Frida Kahlo abgesehen Vom Museo Frida Kahlo gewohnt?	In der Casa de la Cultura	Casa Museo Estudio de Diego Rivera y Frida Kahlo
Wie groß ist der Unterschied zwischen dem Gewicht des Gehirnes eines Mannes und einer Frau?	Das Gehirn eines Mannes wiegt etwa 1,5 Kilogramm, das einer Frau etwa 1,2 Kilogramm	Bei gleicher Statur von Mann und Frau ist das Gehirn bei Männern durchschnittlich 100g schwerer.

Sample Output of German GPT2 1.5B: Plain Wrong

Question	Prediction	Ground Truth
Wer setzte den ägyptischen König 1952 ab?	der ägyptische König Tutanchamun	die Bewegung der Freien Offiziere
Wer ergänzte das Modell eines Kurzzeitgedächtnis?	Der Mathematiker und Philosoph John Locke	das Arbeitsgedächtnismodell von Baddeley
In welcher deutschen Stadt wird der seillose Aufzug getestet?	In der Schweiz, in Deutschland und in Österreich	Rottweil

Interesting papers

- Roberts et al. 2020:
  - larger generative models know more
  - probabilistic vs. fact-based language models
- Aribandi et al. 2022(ExT5): Multi-task learning of QA with Commonsense Reasoning
- Zou et al. 2021: Inverse Prompting for better QA questions
- Zayats et al. 2021, Talmor et al. 2021, Chen et al. 2021: QA over images, tables & text (simultaneously)
- Ram et al. 2021: span selection as pre-training objective for QA
- Longpre et al. 2021: memorization, hallucination, ignoring additional information

References I

- Roberts, A., C. Raffel, and N. Shazeer (2020). "How Much Knowledge Can You Pack into the Parameters of a Language Model?" In: *arXiv preprint arXiv:2002.08910*. arXiv: 2002.08910.
- Chen, W. et al. (Feb. 2021). "Open Question Answering over Tables and Text". In: *arXiv:2010.10439 [cs]*. arXiv: 2010.10439 [cs].
- Longpre, S. et al. (Sept. 2021). "Entity-Based Knowledge Conflicts in Question Answering". In: *arXiv:2109.05052 [cs]*. arXiv: 2109.05052 [cs].
- Ram, O. et al. (2021). "Few-Shot Question Answering by Pretraining Span Selection". In: *arXiv preprint arXiv:2101.00438*. arXiv: 2101.00438.
- Talmor, A. et al. (2021). "Multimodalqa: Complex Question Answering over Text, Tables and Images". In: *arXiv preprint arXiv:2104.06039*. arXiv: 2104.06039.

References II

- Zayats, V., K. Toutanova, and M. Ostendorf (2021). "Representations for Question Answering from Documents with Tables and Text". In: *arXiv preprint arXiv:2101.10573*. arXiv: 2101.10573.
- Zou, X. et al. (2021). "Controllable Generation from Pre-Trained Language Models via Inverse Prompting". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2450–2460.
- Aribandi, V. et al. (Jan. 2022). "ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning". In: *arXiv:2111.10952 [cs]*. arXiv: 2111.10952 [cs].