UNIVERSITÄT
ZU KÖLN

Introduction

HS In Context Learning (ICL) (Summer term 2024)

Nils Reiter,
nils.reiter@uni-koeln.de

April 11, 2024

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Section 1

Kennenlernen

Drei Fragen

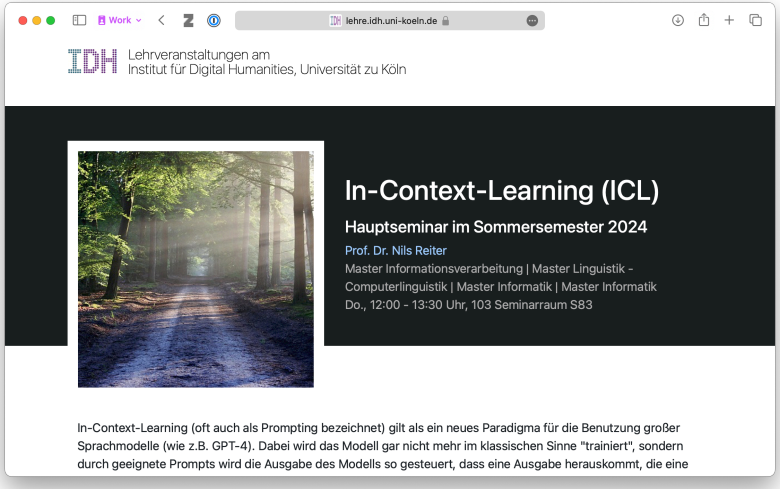
- Wer sind Sie? Worüber haben Sie das letzte Mal gelacht?
- Was stellen Sie sich unter dem Thema “In-Context-Learning” vor?
- Wie geht’s Ihnen?

Section 2

Ablauf und Organisatorisches

Lernziele

- Überblick über das Themenfeld gewinnen
- Tieferes Verständnis einer ICL-Technik erarbeiten
- Aufbereitung eines Themas in Form eines Vortrages



https://uni.koeln/7ZWMG

Module

- MA Informationsverarbeitung: Verarbeitung von Textdaten
- MSc Informatik: Verarbeitung von Textdaten
- MA Linguistik: Profilmodul Computerlinguistik

⚠ Voraussetzungen

Grundkenntnisse im maschinellen Lernen und neuronalen Netzen, Transformer-Architektur sowie der Verwendung großer Sprachmodelle; Programmierkenntnisse in Python.

Modul: Verarbeitung von Textdaten

MA Informationsverarbeitung

Veranstaltung	Kontaktzeit	Selbststudium
Hauptseminar	30	60
Übung	30	60
Kolloquium	30	60
Modulprüfung	–	270

Tabelle: Lehrveranstaltungen im Modul

- 18 Leistungspunkte
- $30\,\%$ der Fachnote

Modul: Profilmodul Computerlinguistik (1C)

MA Linguistik

Veranstaltung	Kontaktzeit	Selbststudium
Hauptseminar	30	60
Projektseminar	30	60
Projektseminar	30	60
Modulprüfung	–	180

Tabelle: Lehrveranstaltungen im Modul

- 15 Leistungspunkte
- $40\,\%$ der Fachnote

Modul: Verarbeitung von Textdaten

M.Sc. Informatik

Veranstaltung	Kontaktzeit	Selbststudium
Hauptseminar	30	240
Übung	30	60
Kolloquium	30	60
Modulprüfung	–	–

Tabelle: Lehrveranstaltungen im Modul

- 15 Leistungspunkte
- $13.1\,\%$ der Fachnote

Modul: Professionalisierung: Forschung (EM 1a)

MA Deutsche Sprache und Literatur

Veranstaltung	Kontaktzeit	Selbststudium
Hauptseminar	30	60
Kolloquium/Oberseminar	30	60
Modulprüfung		180

Tabelle: Lehrveranstaltungen im Modul

- 12 Leistungspunkte
- Das Modul geht nicht in die Fachnote ein.

Studienleistung

Zwei Teile

- Aktive Mitarbeit im Seminar
- Ausarbeitung eines Themas als Seminarsitzung. Schwerpunkt: Experimente, die den Beitrag von ICL zeigen
  - Ablauf: Lehre.IDH

Modulprüfungen

- Thema
  - Findung und Wahl: Ihre Aufgabe
  - Kann, muss aber nicht, etwas mit dem Seminar zu tun haben
  - Mit mir absprechen

Modulprüfungen

- Thema
  - Findung und Wahl: Ihre Aufgabe
  - Kann, muss aber nicht, etwas mit dem Seminar zu tun haben
  - Mit mir absprechen
- Ggf. Praktischer Anteil. Z.B.: Experiment zur automatischen Identifikation eines Textphänomens, Annotationsexperiment, quantitativer Vergleich verschiedener Korpora, ...
- Am Ende: Hausarbeit (wissenschaftlicher Text!) von bestimmter Länge
  - InfoVer, Info, Ling: 8 Seiten
  - DSuL: 7 Seiten

To Do

- Ab sofort
  - Mit Themen beschäftigen
  - Nachdenken/Erinnern: Was waren besonders gute Referate die Sie erlebt haben? Warum waren die gut?
- Ab 15.04., 09:00 Uhr: In Ilias Präferenzen für ein Thema und damit einen Termin angeben

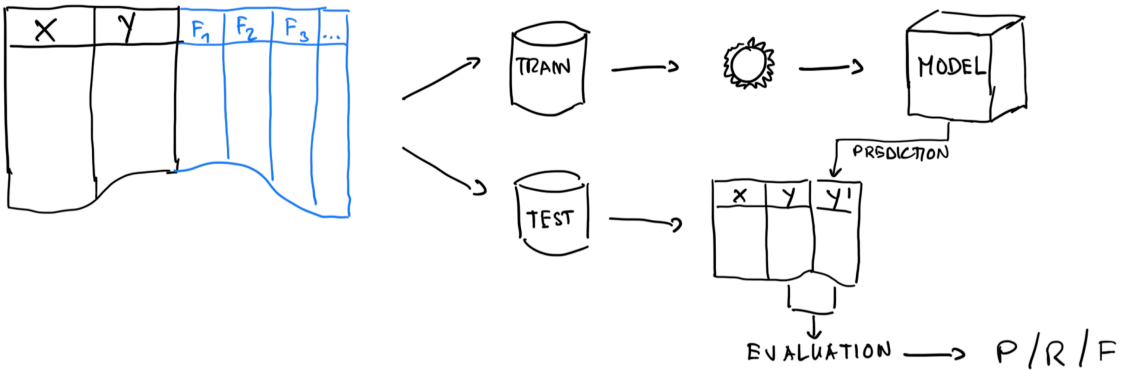
To Do

- Ab sofort
  - Mit Themen beschäftigen
  - Nachdenken/Erinnern: Was waren besonders gute Referate die Sie erlebt haben? Warum waren die gut?
- Ab 15.04., 09:00 Uhr: In Ilias Präferenzen für ein Thema und damit einen Termin angeben
- Nächste Woche
  - Was unterscheidet wissenschaftliche von nicht-wissenschaftlicher Literatur? Wie liest man wissenschaftliche Texte?
  - Wissenschaftlich-technische Themen aufbereiten und präsentieren

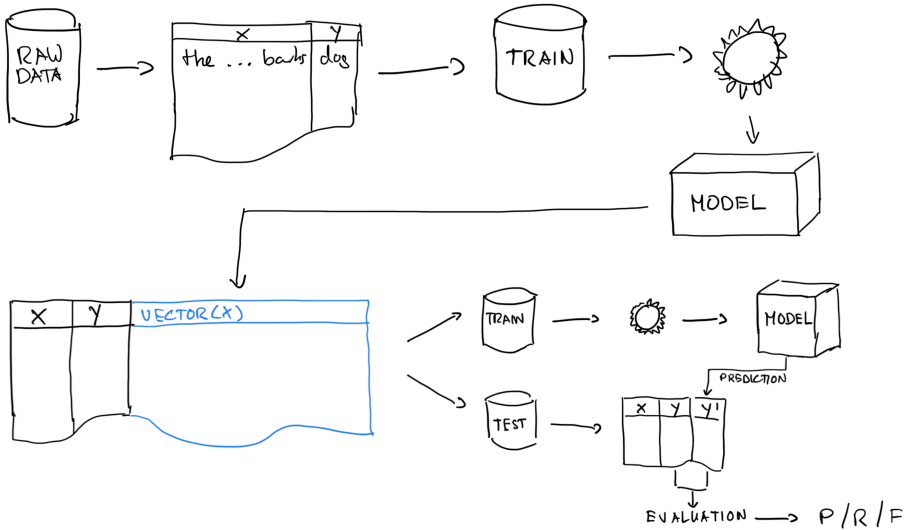
Section 3

Preliminaries

Classical Machine Learning



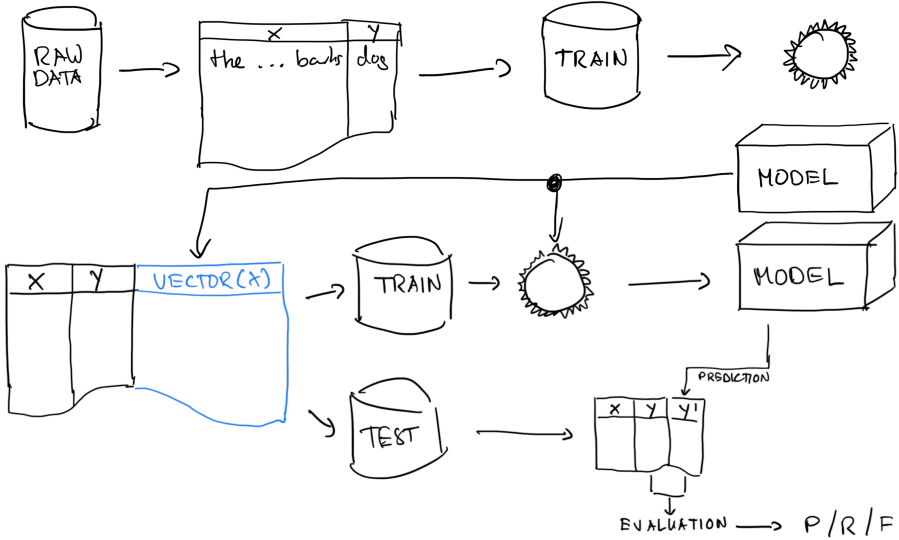
Classical Machine Learning with Learned Representations



BERT

- Transformer-Architecture (i.e., encoder, attention, decoder)
- Trained on
  - Masked language modeling: Cloze
  - Next sentence prediction: Binary classification wether two sentences are actual neighbors
- Usage paradigm
  - Pre-training on huge data sets on generic tasks
  - Fine-tuning for specific task

Transformer



Instruction (Fine) Tuning

- Regular transformer architecture, with specific task: Follow instructions
- Data sets: Manually created or synthetic (often mixed)
- FT data: Triples with instruction, context, expected output

Instruction (Fine) Tuning

- Regular transformer architecture, with specific task: Follow instructions
- Data sets: Manually created or synthetic (often mixed)
- FT data: Triples with instruction, context, expected output

Example (Conover et al., 2023)

- ⚠ No publication except company blog
- 15k examples in 7 areas: creative writing, closed QA, open QA, summarization, information extraction, classification, brainstorming

```
{ "instruction": "What is the currency in use in the Netherlands?",
  "context": "",
  "response": "The currency in use in the Netherlands is the euro.",
  "category": "open_qa" }
```

Using LLMs for Academic Purposes

- Requirements: Reproducibility of experiments, knowledge of influencing factors, tight budget
- Most LLMs are not fully transparent
  - Even the ones that claim to be "open source models"
  - "while there is a fast-growing list of projects billing themselves as 'open source', many inherit undocumented data of dubious legality, few share the all-important instruction-tuning [...], and careful scientific documentation is exceedingly rare" Liesenfeld et al. (2023)
    - 13 features: Open code, LLM data, LLM weights, RLHF data, RLHF weights; License, Code, Architecture, Preprint, Paper, Data sheet; Package, API

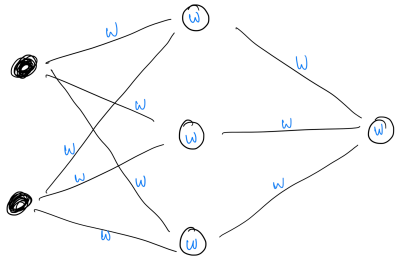
Using LLMs for Academic Purposes

Project	Availability					Documentation					Access methods		
(maker, bases, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture	Preprint	Paper	Data sheet	Package	API
chatGPT	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	-
OpenAI	LLM base: GPT3.5, GPT4			RLHF base: Instruct-GPT								https://chat.openai.com	
StableVicuna-13B	✓	✓	-	-	-	-	-	✓	✓	✗	✗	-	✗
CarperAI	LLM base: LLaMA			RLHF base: oasst1, anthropic						https://huggingface.co/CarperAI/stable-vicuna-13b-delta			
text-generation-webu	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗
oobabooga	LLM base: various			RLHF base: various						https://github.com/Akegarasu/ChatGLM-webui			
MPT-7B-Instruct	✓	✗	✓	-	✗	✓	✓	-	✗	✗	✗	✓	✗
MosaicML	LLM base: MosaicML			RLHF base: dolly, anthropic						https://github.com/mosaicml/llm-foundry#mpt			
Falcon-40B-Instruct	✓	-	✓	-	✓	✓	-	-	-	✗	-	-	✗
TII	LLM base: Falcon 40B			RLHF base: Baize (synthetic)						https://huggingface.co/tiiuae/falcon-40b-instruct			
minChatGPT	✓	✓	✓	-	✗	✓	✓	-	✗	✗	✗	✗	✓
ethanyanjiali	LLM base: GPT2			RLHF base: anthropic						https://github.com/ethanyanjiali/minChatGPT			
trlx	✓	✓	✓	-	✗	✓	✓	-	✗	✗	✗	-	✓
carperai	LLM base: various (pythia, flan, OPT)			RLHF base: various						https://github.com/carperai/trlx			
stanford_alpaca	✓	✓	-	-	✗	-	✓	✓	✗	✗	-	✗	✗
Tatsu labs	LLM base: LLaMA			RLHF base: Self-Instruct (synthetic)						https://github.com/tatsu-lab/stanford_alpaca			
Cerebras-GPT-111M	✓	✓	✓	✓	✗	✓	✓	✓	-	✗	✗	✗	✗
Cerebras, Schramm	LLM base: not open			RLHF base: alpaca (synthetic)						https://huggingface.co/SebastianSchramm/Cerebras-GPT-111M-instruction			
OpenChatKit	✓	✓	✓	✓	✓	✓	✓	✗	-	✗	✗	✓	✗
togethercomputer	LLM base: EleutherAI pythia			RLHF base: OIG						https://github.com/togethercomputer/OpenChatKit			
dolly	✓	✓	✓	-	✗	✓	✓	✓	-	✗	✗	✓	✗
databrickslabs	LLM base: EleutherAI pythia			RLHF base: databricks-dolly-15k						https://github.com/databrickslabs/dolly			
CharRWKV	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
BlinkDL	LLM base: RWKV-LM (own)			RLHF base: alpaca, shareGPT (synthetic)						https://github.com/BlinkDL/ChatRWKV			
BELLE	✓	✓	-	✓	✓	✓	✓	✓	✓	✗	-	✗	✗
LianjiaTech	LLM base: LLaMA, BLOOMZ			RLHF base: alpaca, shareGPT (synthetic)						https://github.com/LianjiaTech/BELLE			
Open-Assistant	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✓	✓
LAION-AI	LLM base: oasst1 (own)			RLHF base: OIG						https://github.com/LAION-AI/Open-Assistant			
xmtf	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RLHF base: xP3						https://github.com/bigscience-workshop/xmtf			

Using LLMs for Academic Purposes

- Requirements: Reproducibility of experiments, knowledge of influencing factors, tight budget
- Most LLMs are not fully transparent
  - Even the ones that claim to be "open source models"
  - "while there is a fast-growing list of projects billing themselves as 'open source', many inherit undocumented data of dubious legality, few share the all-important instruction-tuning [...], and careful scientific documentation is exceedingly rare" Liesenfeld et al. (2023)
    - 13 features: Open code, LLM data, LLM weights, RLHF data, RLHF weights; License, Code, Architecture, Preprint, Paper, Data sheet; Package, API
- This is a problem Balloccu et al. (2024)
  - GPT-3.5 and GPT-4 "have been globally exposed to $\sim 4.7M$ samples from 263 benchmarks"
  - Performance scores of GPT* are likely result of overfitting
  - More on LLM evaluation: May 2nd

Model Sizes



- Number of parameters
  - Simple neural network: 13 parameters
  - More parameters require more training data
- Model sizes (and other details) are increasingly kept secret

Models

Date	Model	Parameters	Training tokens
06/2018	GPT 1	117M	
10/2018	BERT (large)	340M	3.3B
02/2019	GPT 2	1.5B	10B
10/2019	T5	11B	34B
05/2020	GPT 3	175B	300B
04/2022	PaLM	540B	768B
07/2022	BLOOM	175B	350B
02/2023	LlaMA	65B	1.4T
03/2023	GPT-4		

Table: LLM stats, according to Wikipedia. $1M = 10^6, 1B = 10^9 = 1$ Milliarde (de), $1T = 10^{12}$

Cost

Training

- All concrete numbers are estimations
  - I.e., no one who knows has actually confirmed a price tag
- Development cost are much higher than training the model once
- PaLM (530B params, 1.6T corpus): $9\,\text{M\$}$ to $23\,\text{M\$}$ Heim (2022)
- MosaicML GPT-70B (70B params, 1.4T corpus): $2.5\,\text{M\$}$ Venigalla/Li (2022)

Cost

Training

- All concrete numbers are estimations
  - I.e., no one who knows has actually confirmed a price tag
- Development cost are much higher than training the model once
- PaLM (530B params, 1.6T corpus): $9\,\text{M\$}$ to $23\,\text{M\$}$ Heim (2022)
- MosaicML GPT-70B (70B params, 1.4T corpus): $2.5\,\text{M\$}$ Venigalla/Li (2022)

Using (= "inference")

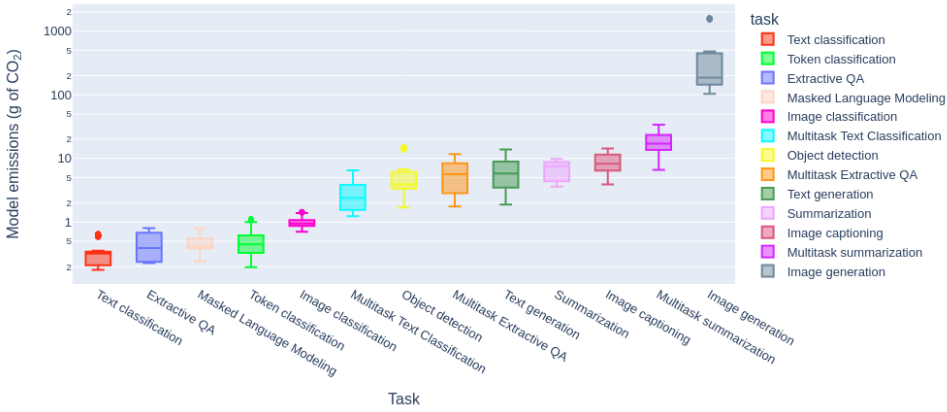
- GPT-4 Turbo: $10\,\$/1\text{M}$ input tokens $+$ $30\,\$/1\text{M}$ output tokens OpenAI Inc. (2024)

Cost

- LLMs require significant amounts of power
- Power production often involves burning fossil fuels, which emits CO_2 (or equivalents)
- Rising CO_2 costs will impact prices for AI/LLM applications
- Only few studies, developing research area Luccioni et al. (2023); Strubell et al. (2019)

Cost

Luccioni et al. (2023)



Cost

Luccioni et al. (2023)

	BLOOMz-7B	BLOOMz-3B	BLOOMz-1B	BLOOMz-560M
Training energy (kWh)	51,686	25,634	17,052	10,505
Finetuning energy (kWh)	7,571	3,242	1,081	543
Inference energy (kWh)	1.0×10^{-4}	7.3×10^{-5}	6.2×10^{-5}	5.4×10^{-5}
Cost parity (# inferences)	592,570,000	395,602,740	292,467,741	204,592,592

Table: Cost parity: Number of inferences required to sum to the training cost.

Form vs. Meaning

- Do LLMs know and understand things, or are they 'just' simulating?
- Rhetorics: Which metaphors do we use, in particular in communication to the public?
  - Are we talking about "artificial intelligence"?

Form vs. Meaning

- Do LLMs know and understand things, or are they 'just' simulating?
- Rhetorics: Which metaphors do we use, in particular in communication to the public?
  - Are we talking about "artificial intelligence"?
- LLMs and "natural language understanding" (NLU)

  bender_climbing_2020; Bender et al. (2021)
  - LLMs only ever learn about linguistic surfaces, there is nothing else
  - But language involves mapping from form to meaning
  - Even if they can reply like a human, they cannot understand as a human
    - (also an argument against the significance of the Turing test)
  - Supporting evidence: LLMs are sensitive to input variation

Form vs. Meaning

Stochastic Parrots (Bender et al., 2021)

- Communication is a jointly constructed activity
- Rests on shared common ground, mutual awareness of this sharing, and communicative intents
- "if one side of the communication does not have meaning, then the comprehension of the implicit meaning is an illusion arising from our singular human understanding of language (independent of the model)" Bender et al. (2021, 616)

Section 3

Preliminaries

Summary

- LLMs: State of the art technology in NLP
- Many open issues, not all of them technical
- Academic study of LLMs is possible, but one has to be careful
  - Ideally: Local control with replicable experimental conditions
- Training/using is expensive in multiple ways
- It's not clear what kind of meaning LLMs learn, if any
  - Beware of your metaphors when speaking about LLMs

To Do

- Ab 15.04., 09:00 Uhr: In Ilias Präferenzen für ein Thema und damit einen Termin angeben
- Was waren besonders gute Referate die Sie erlebt haben? Warum waren die gut?

References I

- Balloccu, Simone/Patrícia Schmidtová/Mateusz Lango/Ondrej Dusek (2024). "Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham/Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 67–93. URL: https://aclanthology.org/2024.eacl-long.5.
- Bender, Emily M./Timnit Gebru/Angelina McMillan-Major/Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. event-place: Virtual Event, Canada. New York, NY, USA: Association for Computing Machinery, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

References II

- Conover, Mike/Matt Hayes/Ankit Mathur/Jianwei Xie/Jun Wan/Sam Shah/Ali Ghodsi/Patrick Wendell/Matei Zaharia/Reynold Xin (2023). *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. URL: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm (visited on 06/30/2023).
- Heim, Lennart (2022). *Estimating PaLM's training cost*. URL: https://blog.heim.xyz/palm-training-cost/.
- Liesenfeld, Andreas/Alianda Lopez/Mark Dingemanse (2023). "Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators". In: *Proceedings of the 5th International Conference on Conversational User Interfaces*. CUI '23. event-place: , Eindhoven, Netherlands, New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3571884.3604316. URL: https://doi.org/10.1145/3571884.3604316.

References III

- Luccioni, Alexandra Sasha/Yacine Jernite/Emma Strubell (2023). *Power Hungry Processing: Watts Driving the Cost of AI Deployment?* DOI: 10.48550/arXiv.2311.16863. URL: https://arxiv.org/abs/2311.16863.
- OpenAI Inc. (2024). *Pricing*. URL: https://openai.com/pricing.
- Strubell, Emma/Ananya Ganesh/Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen/David Traum/Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: https://aclanthology.org/P19-1355.
- Venigalla, Abhi/Linden Li (2022). *Mosaic LLMs: GPT-3 quality for <$500k*. URL: https://www.databricks.com/blog/gpt-3-quality-for-500k.