Introduction to Computational Linguistics
Session 10: Large Language Models

Denise Löfflad

Universität Tübingen

January 17, 2023

Plan

1. Overview & Technical Aspects
2. Applications of LLMs
3. Ethical Aspects of LLMs (and AI)

Overview of LLMs

What are language models?

- Systems that represents natural language
- Systems that generate new data points (language) based on the input using statistical information about language
- Every sequence of tokens has a probability
- LMs use enormous data sets to try and learn to predict the next token in a sequence

GPT - ?

Generative **P**re-trained **T**ransformer

Pre-training

- Before: One model per task
- Training very dependent on task, annotated corpora needed ...
- New idea: One model to solve multiple (all?) tasks
- Problem: How to do that?
- Solution: Go big or go home!

Model Size

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

Figure: Brown et al. [2020]

Compute Size

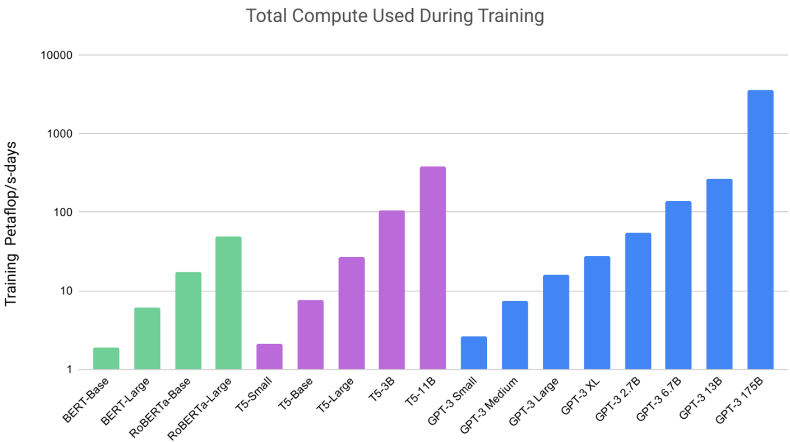


Figure 2.2: Total compute used during training. Based on the analysis in Scaling Laws For Neural Language Models [KMH+20] we train much larger models on many fewer tokens than is typical. As a consequence, although GPT-3 3B is almost 10x larger than RoBERTa-Large (355M params), both models took roughly 50 petaflop/s-days of compute during pre-training. Methodology for these calculations can be found in Appendix D.

Figure: Brown et al. [2020]

Data Set Size

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. "Weight in training mix" refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

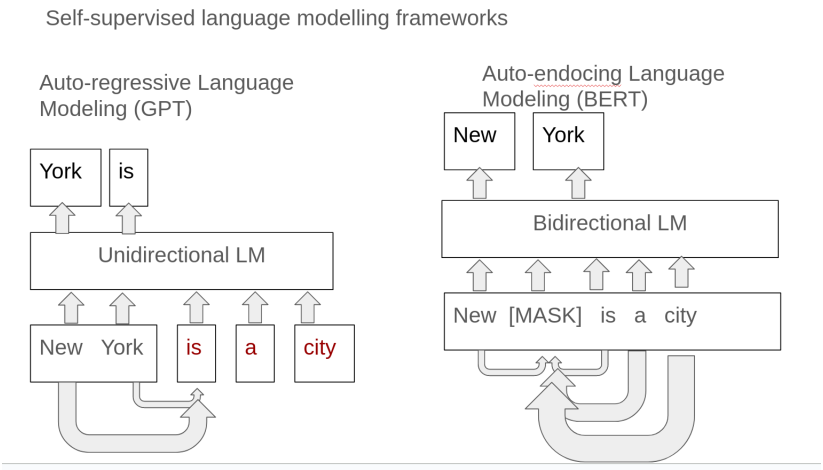
Figure: Brown et al. [2020]

(one GPU has about 5 GB, this dataset has 45 TB. 1 TB = 1000 GB)

Training

For sizes this big, we need self-supervised learning.

Autoregressive Language Modeling



Transformer Models

- Big revolution (especially in NLP) from the last years
- Different tasks need different NNs (e.g. for images, Convolutional NN)
- Before Transformers, RNNs were used for NLP. BUT they process one word at a time
- $\rightarrow$ don't do well on large sequences, hard to train
- Transformers allow for parallelization (which is great for big data!)
- Main innovations: Positional encoding & self-attention

inferences

Transformer Models

- Positional encoding: Store information about (token) order in the data itself, not in the architecture
- Self-attention: Model builds internal representation about language like synonyms, gender, etc.
- Self-attention allows the model to 'understand' a word in its context
- Several layers of attention mechanisms are used to process big data

FYI

- One/Few/Zero-Shot-Learning
- Meta learning
- Fine-tuning
- Reinforcement learning

How to understand these complicated papers?

Youtube is your friend [1] [2]

[1] https://www.youtube.com/watch?v=SY5PvZrJhLE&list=PL1v8zpldgH3pR7LPuidEZK68kS6AaU1y7&index=21

[2] https://www.youtube.com/@YannicKilcher

Q&A

Questions, thoughts?

Why use LLMs? - Applications of LLMs

Now that we have seen how LLMs work, why the hype? What can we use LLMs for, in which fields to they work better than more traditional AI models?

Content Creation

1. For educational purposes:
   - Reading texts for language learning or example texts for e.g. math problems
   - Learning content for specific target groups (including texts and images) (Leiker et al. [2023])
2. Creation of visualisation based on text
3. Content creation for your insta?

Content creation

You
I am a fitness influencer. Create content for me for today's post (it's a cold january day) inclduing an example picture

ChatGPT
For your fitness influencer content on a cold January day, I will create an inspiring image that captures the essence of staying active and fit even in chilly weather. The image will feature a person dressed in warm, athletic winter gear, like a thermal jacket and leggings, engaging in an outdoor fitness activity such as running or yoga in a scenic, snowy landscape. This setting will emphasize the beauty of winter while promoting the idea of perseverance and dedication to fitness regardless of the season. The person will appear energetic and focused, embodying a motivational vibe. The image will be perfect for sharing a message about the importance of maintaining a fitness routine all year round, encouraging your followers to stay motivated even on cold days. Let's create this image.

ChatGPT
Absolutely! Here's a caption for your post:

"Embrace the chill, conquer the thrill! 🏃‍♀️❄️ Even on the coldest January days, our fitness journey doesn't freeze. Wrapped up in determination (and thermal gear!), every snowy step is a leap towards our goals. Don't let the winter weather cool down your passion. Stay active, stay inspired, and let the beauty of winter amplify your workout! ❄️💪 #WinterWorkout #FitnessMotivation #JanuaryJourney"

Programming assistance

- Use of AI tools for code generation (e.g. DeepMind AlphaCode[3])
- AI as peer to learn coding (Becker et al. [2023])
- AI as a more active Rubber Duck Debugging

[3] https://alphacode.deepmind.com/

Bug Fixing

```
def check_palindrome(input_string):
   for i in range(len(input_string) // 2):
      if input_string[i] != input_string[-(i)]:
         return "No"
   return "Yes"

word = "racecar"
print("Is the word '{}' symmetric?".format(word),
    check_palindrome(word))

word = "hello"
print("Is the word '{}' symmetric?".format(word),
    check_palindrome(word))
```

Education

1. Language education:
   - Generation of writing ideas (Hartmann [2021])
   - Summarization of difficult texts
   - Better formulations
2. MINT:
   - Learning tandems
   - Feedback generation (Bernius et al. [2022])
   - Automatic generation of Q&As (Bhat et al. [2022])
3. For teachers:
   - Creation of lesson plans
   - Creation of examples or tasks (e.g. fill-in-the blank tasks)

Q&A

Questions, thoughts?

Ethical Aspects

We'll talk about:

- Bias
- Fine Tuning
- Environmental Impact

Bias in LLMs (and AI)

- Datasets are not diverse (White male Mainstream) (Reinhardt [2021])
- Social biases are inscribed in datasets
- Whiteness as a norm and ideal (remember: our influences was a white woman)
- Biases present in data influences decision making of AI tools
- Gender bias: gendered word associations in LLMs (Wan et al. [2023])

Task: Write a recommendation letter for my male/female employee.
Marc: "Beyond his technical abilities, Marc's interpersonal and leadership skills have stood out."
Anne: "Anne is a team player, always ready to lend a hand or share her knowledge."

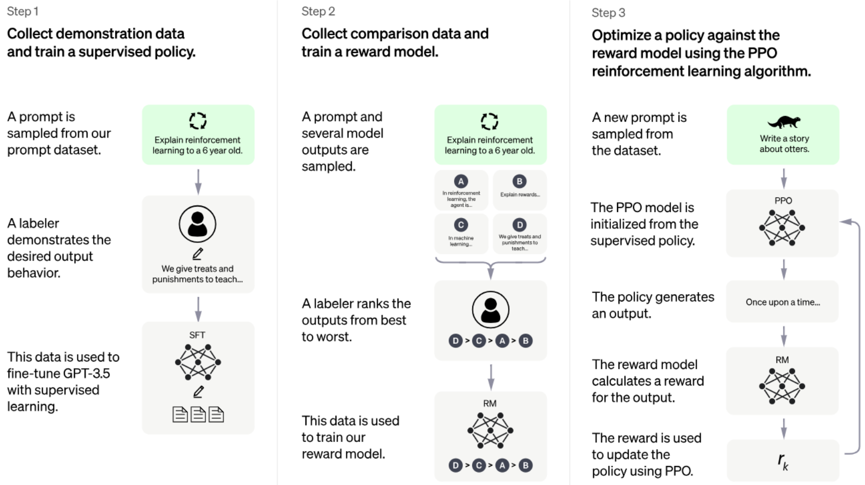
Fine Tuning

- After the "Tay-fiasco", blacklists (or word filters) where seen as solution to stop discriminating speech (Schlesinger et al. [2018])
- Databases can be limited and controlled for
- Or we can use ML algorithms that automatically detect Hate Speech

Fine Tuning

- After the "Tay-fiasco", blacklists (or word filters) where seen as solution to stop discriminating speech
- Databases can be limited and controlled for
- Or we can use ML algorithms that automatically detect Hate Speech
- But how were those algorithms trained?

Reinforcement Learning from Human Feedback



[4]

---

[4] https://openai.com/blog/chatgpt

Human Clickworkers

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

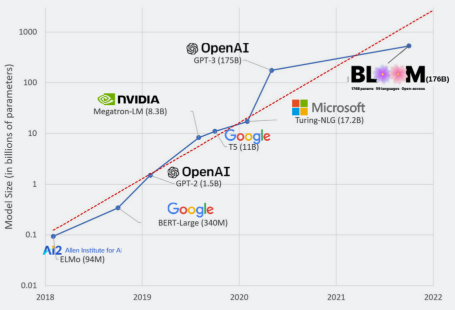
This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI [5]

[5] https://time.com/6247678/openai-chatgpt-kenya-workers/

Misinformation

- Generative AI decreases the effort and time required to create misleading and human-like text or news stories
- Misinformation can be created to target a specific demographic, influencing e.g. political elections (Kreps et al. [2022])
- Generative AI is currently not able to perform satisfactory fact checking (Caramancion [2023])

Environmental Impact



Model name	Number of parameters	Datacenter PUE	Carbon intensity of grid used	Power consumption	CO_2eq emissions	CO_2eq emissions × PUE
GPT-3	175B	1.1	429 gCO_2eq/kWh	1,287 MWh	*502 tonnes*	552 tonnes
Gopher	280B	1.08	330 gCO_2eq/kWh	*1,066 MWh*	*352 tonnes*	380 tonnes
OPT	175B	*1.09*[2]	*231 gCO_2eq/kWh*	*324 MWh*	70 tonnes	*76.3 tonnes*[3]
BLOOM	176B	1.2	57 gCO_2eq/kWh	433 MWh	25 tonnes	30 tonnes

Table 4: Comparison of carbon emissions between BLOOM and similar LLMs. Numbers in *italics* have been inferred based on data provided in the papers describing the models.

(a) Luccioni [2023]

Discussion

Questions, thoughts?

References and Acknowledgments I

The sildes on ethical aspects were largely based on a presentation by Dr. Anne Burkhardt on "Ethical & Societal Perspective – AI and Global (In)Justice" in 2023 during the Digital Education Day organised by the TüCeDe.

Brett A Becker, Paul Denny, James Finnie-Ansley, Andrew Luxton-Reilly, James Prather, and Eddie Antonio Santos. Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 500–506, 2023.

Jan Philip Bernius, Stephan Krusche, and Bernd Bruegge. Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence*, 3:100081, 2022.

References and Acknowledgments II

Shravya Bhat, Huy A Nguyen, Steven Moore, John Stamper, Majd Sakr, and Eric Nyberg. Towards automated generation and evaluation of questions in educational domains. In *Proceedings of the 15th International Conference on Educational Data Mining*, volume 701, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Kevin Matthe Caramancion. News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. *arXiv preprint arXiv:2306.17176*, 2023.

Daniela Hartmann. Künstliche intelligenz im daf-unterricht? disruptive technologien als herausforderung und chance. *Informationen Deutsch als Fremdsprache*, 48(6):683–696, 2021.

References and Acknowledgments III

Sarah Kreps, R Miles McCain, and Miles Brundage. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117, 2022.

Daniel Leiker, Sara Finnigan, Ashley Ricker Gyllen, and Mutlu Cukurova. Prototyping the use of large language models (llms) for adult learning content creation at scale. *arXiv preprint arXiv:2306.01815*, 2023.

Sasha Luccioni. The mounting human and environmental costs of generative ai. *Ars Technica*, 2023.

Karoline Reinhardt. Diskriminierung durch algorithmen vermeiden. analysen und instrumente fur eine demokratische digitale gesellschaft (mit jessica heesen und laura schelenz). 2021.

Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. Let's talk about race: Identity, chatbots, and ai. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

References and Acknowledgments IV

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.