Large Language Models

Jindřich Libovický

April 24, 2024

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Outline

Decoder-only models

Model scaling

Introduction-base Fine-tuning

Societal impact of LLMs

Decoder-only models

Reminder: General Architecture Overview

1.

Embed input words.

Get a sequence of continuous vectors

2.

Contextualize input.

Apply a sequence processing architecture and get contextual representation.

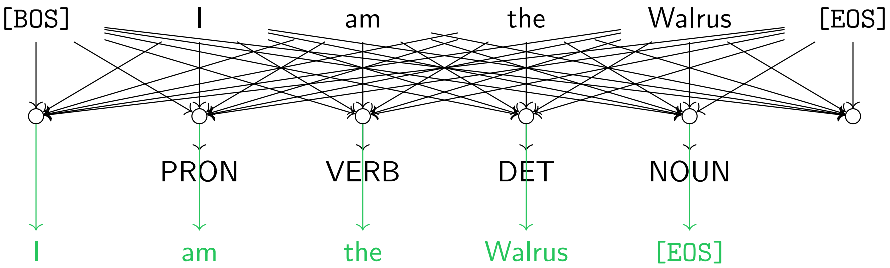
3.

Get some output.

Typically classification or labeling.

From Sequence Labeling to Decoding

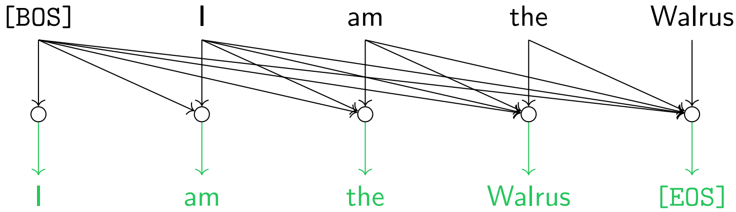
Transformers for sequence labeling



What if we labeled the sequence with what the next word is?

LM as Sequence Labeling

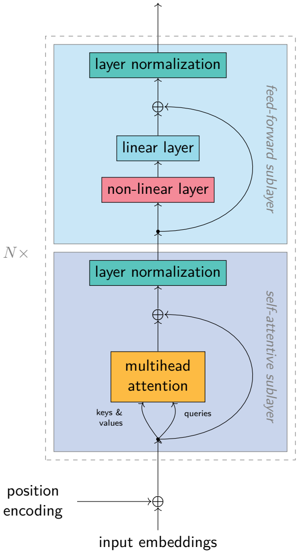
We need to modify the attention **not to attend the right context**



LMs estimate probability of a text: P("I am the Walrus") =

P("I"|[BOS]) · P("am"|[BOS] I) · P("the"|[BOS] I am) · P("Walrus"|[BOS] I am the)·
P("[EOS]" | [BOS] I am the Walrus)

Reminder: Transformer Architecture



- Several layers (original paper 6)
- Each layer 2 sub-layers: self-attention and feed-forward layer
- Everything inter-connected with residual connections

Feedforward-layer

$$F(X) = W_2\,\mathrm{ReLu}\left(W_1 X + b_1\right) + b_2$$

Reminder: Multi-head scaled dot-product attention

Scaled dot-production attention

$Q = (q_1, \ldots, q_n)$: queries, K: keys, V: values

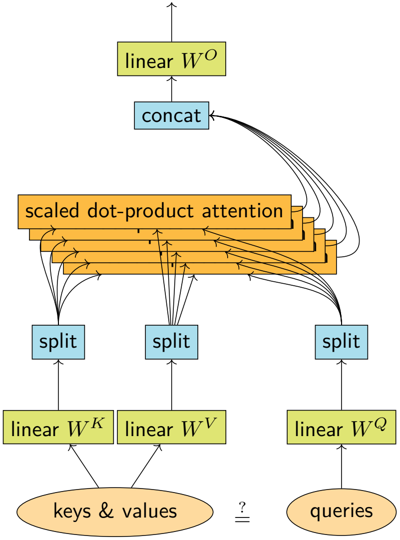
$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\overbrace{\left(\frac{QK^\top}{\sqrt{d}}\right)}^{\text{similarity matrix}} V$$

Multi-head setup

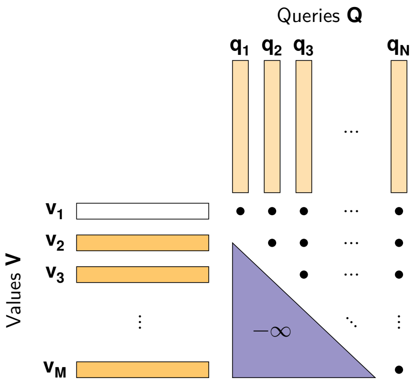
$$\mathrm{Multihead}(Q, V) = \overbrace{(H_1 \oplus \cdots \oplus H_h)}^{\text{concatenate head outputs}} W^O$$

$$H_i = \mathrm{Attn}(QW_i^Q, VW_i^K, VW_i^V)$$

W_i^Q, W_i^K, W_iV head-specific projections



Triangular Mask to Make Training Work



- Target is known at training: don't need to wait until it's generated
- Self attention can be parallelized via matrix multiplication
- Prevent attentding the future using a mask

From Probability to Generation

LM itself only **computes probability**

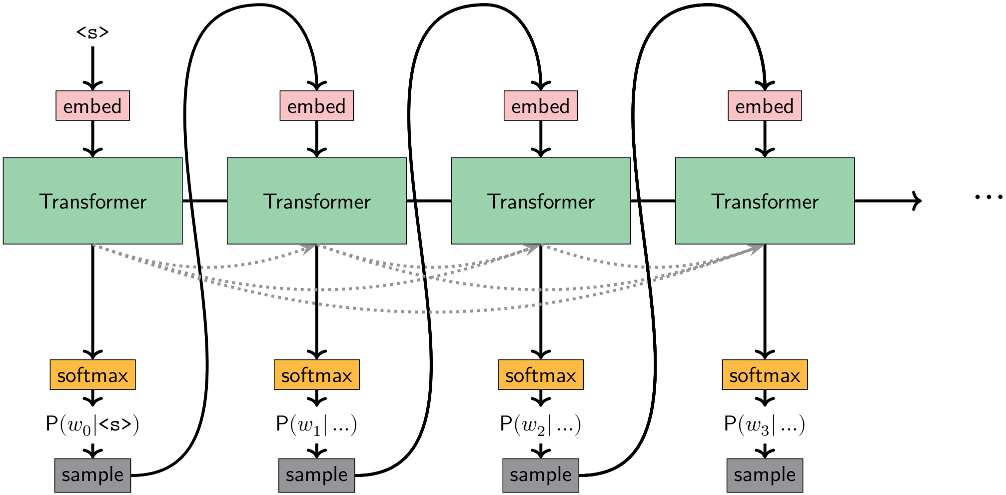
$$\mathsf{P}(w_{n+1} | w_0, \ldots, w_n)$$

We need an **inference algorithm**.

- In machine translation: beam search to search for maximum probable target
- With LMs random sampling — problem with **exposure bias**
  I.e., LM would behave strange if it sampled something inprobable
- Solution: **top-**k sampling or **nucleus** sampling (= top x probability mass)

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`

Inference



<s>
embed
embed
embed
embed
Transformer
Transformer
Transformer
Transformer
...
softmax
softmax
softmax
softmax
P(w_0|<s>)
P(w_1| ...)
P(w_2| ...)
P(w_3| ...)
sample
sample
sample
sample

Model scaling

Timeline 2017–2018

06/2017 Transformer Architecture

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017

02/2018 ELMo – LM pre-training with RNNs for finetuning: 93.6M params.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL `https://aclanthology.org/N18-1202`

06/2018 GPT-1 – Generative LM, in this scale for finetuning only, worse than *BERT*

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018

Timeline 2017–2018

10/2018 BERT – LM pretraining but with Transformers: 345M params.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`

02/2019 GPT-2 – generative model 1.5B parameters

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019

04/2020 GPT-3 – 175B parameters

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

12/2022 ChatGPT

Few-shot/In-context learning

The gray text is **model input**, the black text is **continuation**.

```
Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you
requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you
requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.
```

```
Poor English input: Please provide me with a short brief of the design you're looking for and
that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're
looking for and that would be nice if you could share some examples or projects you have
done before.
```

```
Poor English input: The patient was died.
Good English output: The patient died.
```

Source GPT-3 Paper, Brown et al., 2020, Fig 3.17

- No update of the parameters
- The LM just continues in the text in the same style

Zero-shot learning and Prompting

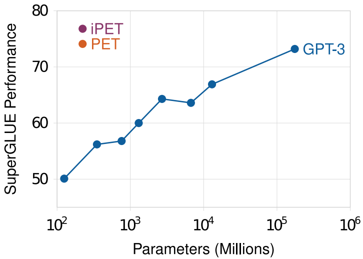
Few shot learning success depend on how the **task is formulated**.
(e.g., number of few-shot, examples needed, task description might be enough)

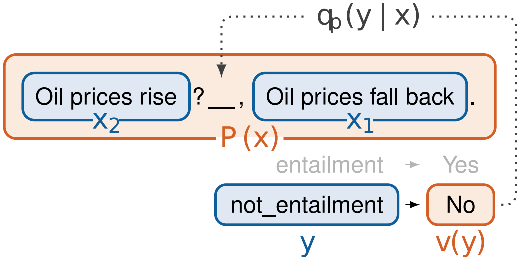
$$\Downarrow$$

Art of finding the correct prompt = **Prompt Engineering**

Later discovered that smaller LMs are also few-shot learners...

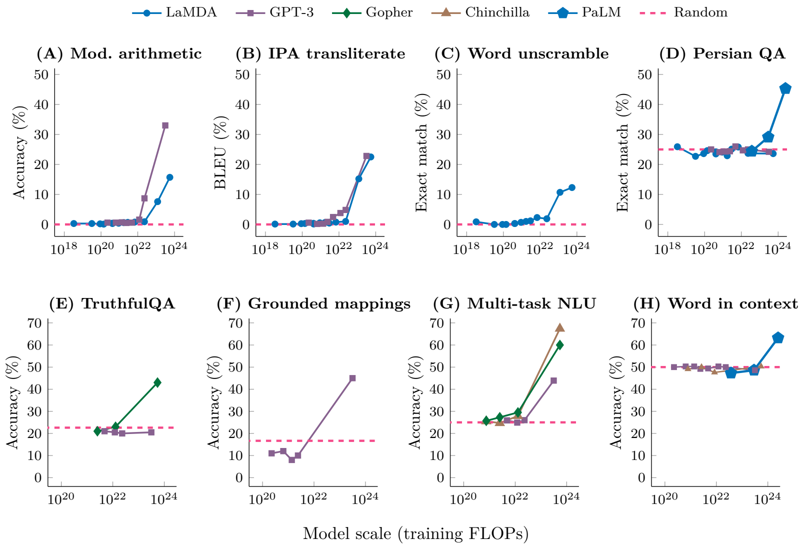
Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.185. URL `https://aclanthology.org/2021.naacl-main.185`, below Figures 1 and 2.





Just replace in-context learning with clever finetuning, only consider predefined set of answers.

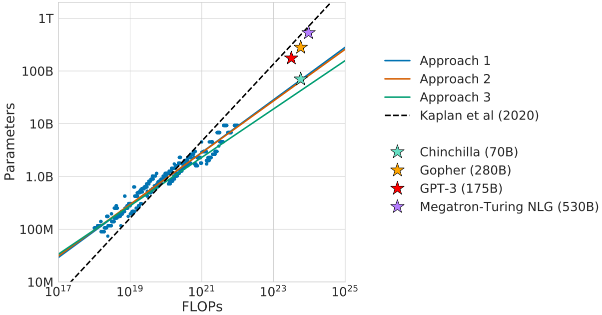
Emergent Properties of LMs (2)



Source: Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022

DeepMind's Chinchila experiments: longer traning can compensate for parameter count.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022, Figure 1



This trick used in Meta's LLAMA: GPT-3 quality with 30B parameters.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023

Emergent Properties of LMs (2)

- New abilities of LMs **emerge with size** – must be **discovered**
- Counterargument: Retrospectively with a **continuous metric**, everything is **gradual**

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023

Introduction-base Fine-tuning

From LM to an Assistant

Finding the right prompt for a given task is alchemy...

What if we **finetuned** the LM for **instruction-like prompts**?

$\Rightarrow$ Finetuning on instructions, reinforcement learning with human feedback
Started with InstructGPT, ChatGPT, ...

Following X slides are based on Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022

Three steps of InstructGPT

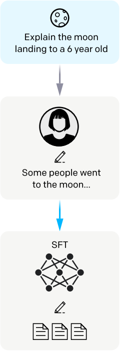
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.



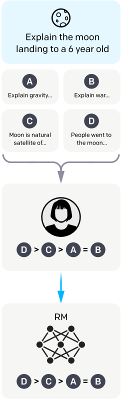
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

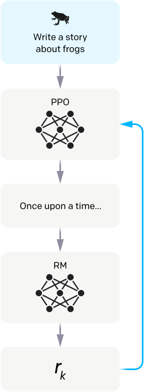
Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

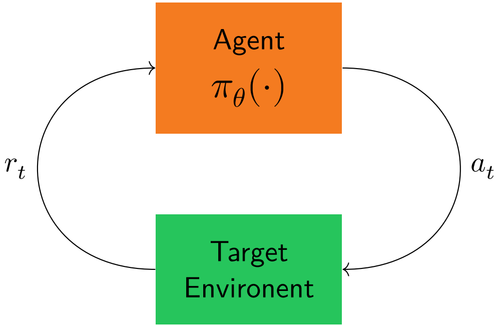


The InstructGPT paper, Ouyang et al., 2022, Figure 2

Supervised Finetuning

- **Annotators write scripts** of conversation with the assistant
- Scripts are used for **direct finenuting**
- 10^5–10^6 conversations are needed in this stage

Reinforcement Learning



- Agent with parameters θ, gets a distribution over possible actions π_θ
- $a_t \cdots$ Action taken in step t
- $r_t \cdots$ Reward in step t

Problem with training: We **cannot compute gradient** $\partial r_t / \partial \pi_\theta$ because we **sample a single action**.

Actions & Reward

Actions

$$a_t$$

Words sampled from the model.

I.e., the inference algorithm is a part of the training.

Reward

$$r_t$$

Simulated human feedback.

A neural network prediction how would human annotators like the output.

Simulating Human Feedback

- Sample multiple K answers for a prompt
- Annotators rank them $\Rightarrow \binom{K}{2}$ pair-wise comparions
- Fine-tune BERT-like model to predict the comparison

Loss function over dataset of prompts x:

$$-\sum_{\text{prompt } x} \frac{1}{\binom{K}{2}} \sum_{\substack{y_w, y_l \\ w \text{ better than } l}} \log\left[\sigma(r(x, y_w) - r(x, y_l))\right]$$

Learning from the Reward Function (Sketch)

- Goal: Optimize expected reward
- We can approximate the gradient of reward with respect to the action distribution (policy gradient algorithms):

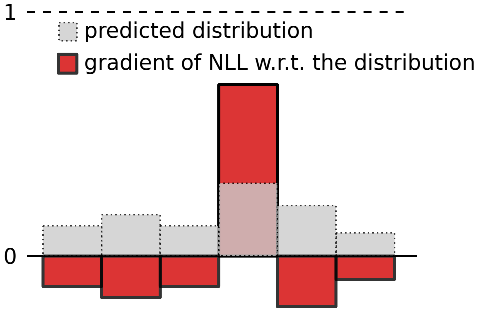
$$\frac{\partial}{\partial \pi_\theta} \mathbb{E}_{a \sim \pi_\theta} r(x, a) \approx \nabla \log \pi_\theta(a|x) \tilde{A}$$

  $\tilde{A}$ is advantage (estimate reward gain $=$ reward shifted by a smart constant)
- Dark magic: Construct a differentiable function that has a gradient like this $+$ some good properties
- InstructGPT uses Proximal Policy Optimization

  John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017
- Additionally: Minimize KL-Divergence from the supervised finetuned model

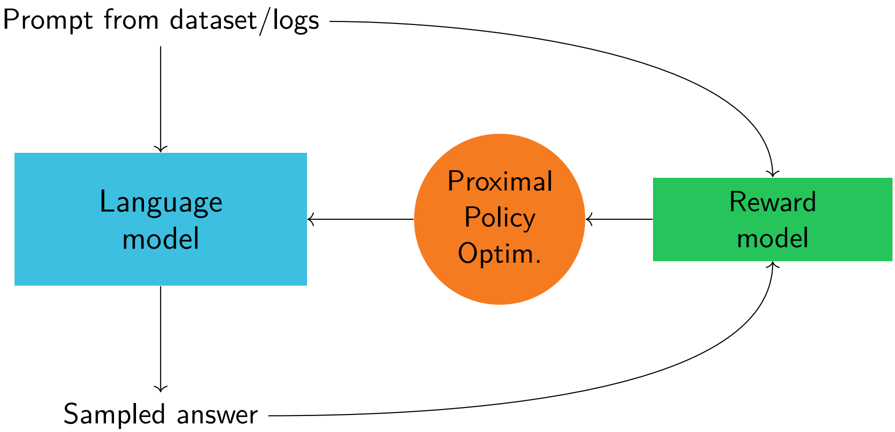
Intuition: Analogy to the Derivative of Cross-Entropy

$$\frac{\partial L(P_y, y^*)}{\partial l} = -\frac{\partial}{\partial l} \log \frac{\exp l_{y^*}}{\sum_j \exp l_j} = \mathbf{1}_{y^*} - P_y(y^*)$$



We want to have similar gradient but with **selected action and reward** (i.e., positive advantage when reward is better than expectation)

Training Overview



It is not only ChatGPT

- Alpaca, Vicuna: Finetuned Meta's LLAMA at Standford
- OpenAssistant: crowd-sourced dataset, models based on LLAMA and Pythia

...in May 2023 for most NLP tasks task-specific models are by large margin better.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*, 2023

Societal impact of LLMs

Problems with LLMs

- Changes in labour productivity will have economical consequences
- Labour productivity increases only somewhere:
  Only available in languages of Global North
- Biases against already underprivileged social groups

Training Data Might be Problematic

Crawling the Internet — not representative, people with extreme/wird opinions write more texts than the rest of society

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021

Crowd-sourcing — using cheep labour, so-called gig economy – precarization of labour

Mary L Gray and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books, 2019

Mining existing databases — unpaid labour, nontransparent "payment" for "free services"

Nick Couldry and Ulises A Mejias. *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press, 2020

LLMs are Stochastic Parrots

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021

The most **famous criticism** of LLMs (already from 2021)

- **Carbon footprint**: Different people benefit from the technology and different people carry the consequences
- Documented examples as **gender and racial bias** (in dubious applications in the US)
- **Nontransparent data curation**: authors decides about values in the text without much outside control

Narrated as a consequence of LMs **not capturing meaning**.

Octopus: Philosophical Background for the Parrots

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021

Assumption: Meaning is a (mathematical) relation on $M \subset I \times E$
$e \in E$ expression, $i \in I$ intent

Thought experiment: An octopus listens to human conversation over line and then simulates human conversation

- It never saw the world above the sea
- In only sees the expressions, no clue what intentions might be about

$\Rightarrow$ no way of understanding this type of meaning

Without meaning, (ethical) reasoning is impossible $\Rightarrow$ LMs inherently harmful

Arguments against the Octopus

- Alternative **established theories of meaning** that dot not have this problem
- Empirical evidence rebute the main points (beliefs and communication intentes can be identified via probing)

Jacob Andreas. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.findings-emnlp.423`

Summary

1. Generating text = sampling from a language model
2. Scaling causes LMs to gain more capabilites
3. Instruction following is learned via reinforcement learning
4. Ethical problems: Reinforcing already existing societal issues