

# Large language models: What does "large" and "language" mean here?

Jindřich Libovický

15. 9. 2023



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



unless otherwise stated

# Outline

1. Language models and neural networks
2. Selection of language model types
3. NLP Tasks solved with language models
4. Generative models
5. From LMs to Assistants
6. Stochastic parrots and other problems
7. Research at ÚFAL

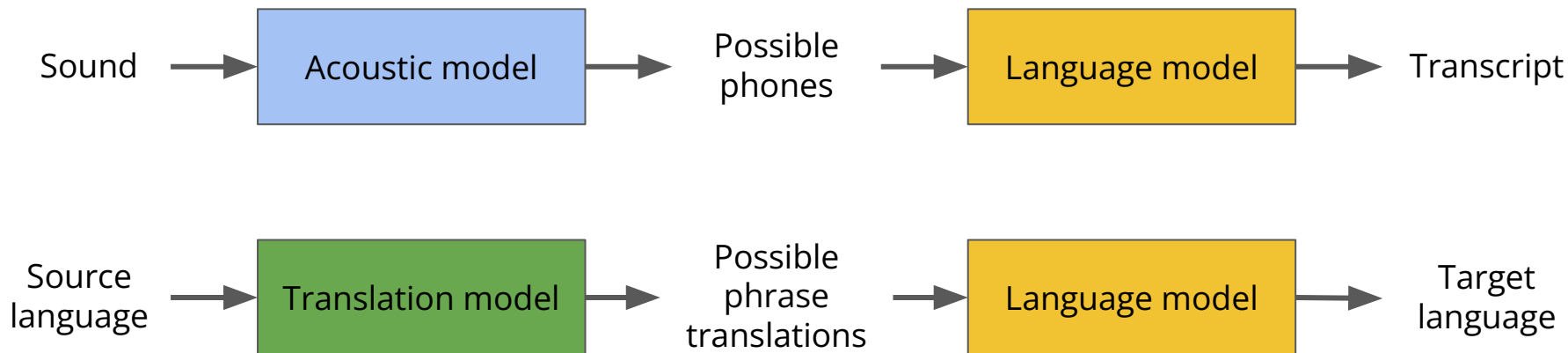
# How do LMs work

**Estimate the probability of a  
word / sentence / text  
in a context.**

# Where do LMs come from

Since 1990's

An important component in speech recognition and machine translation



## LMs used as...

1. Models of what is good/bad in a language
2. Representation learning models
3. Generative models

# Language Models

Goal: Estimate **probability of text**

$P(\text{she is a doctor}) > P(\text{dog's name is dog}) > P(\text{dsa ds gf afgra fw})$

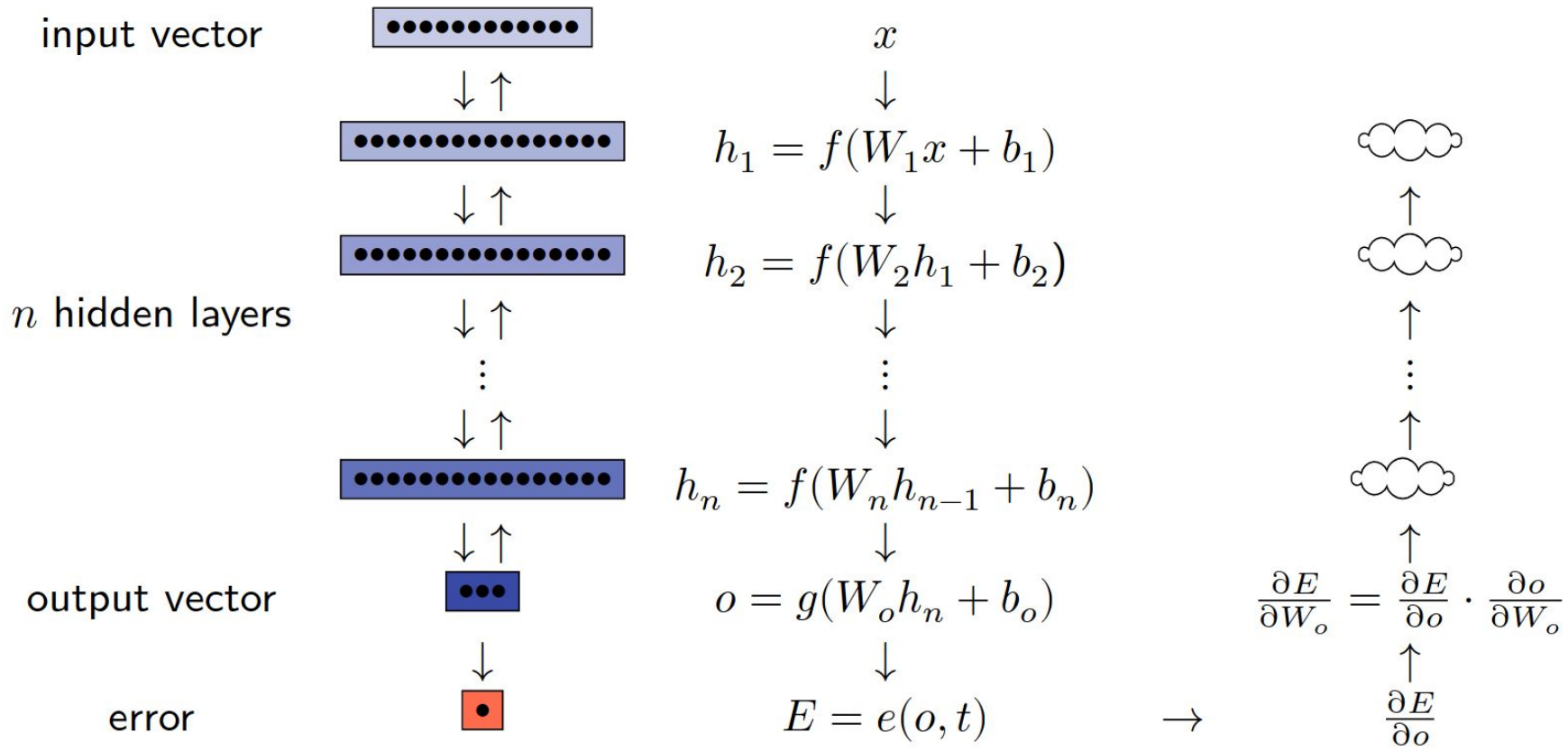
- Historic significance as **a component** in machine translation or speech recognition systems
  - Noisy channel model
  - $P(\text{target} | \text{source}) = P(\text{source} | \text{target}) \times P(\text{target}) / P(\text{source})$
  - best target =  $\text{argmax } P(\text{source} | \text{target}) \times \mathbf{P(\text{target})}$
- Statistical: *trained* to **maximize likelihood** of the training data

# Neural Networks

- Machine learning model
  - Parameterized function mapping the input to a prediction
- Built around non-linear transformations of intermediate results
  - “Layers”
  - Affine transformations followed by non-linear “activation function”
  - Great match with parallel processing of batches of data on GPUs
- Structured architecture
  - Recurrent networks
  - Encoder-decoder
  - Attention mechanism

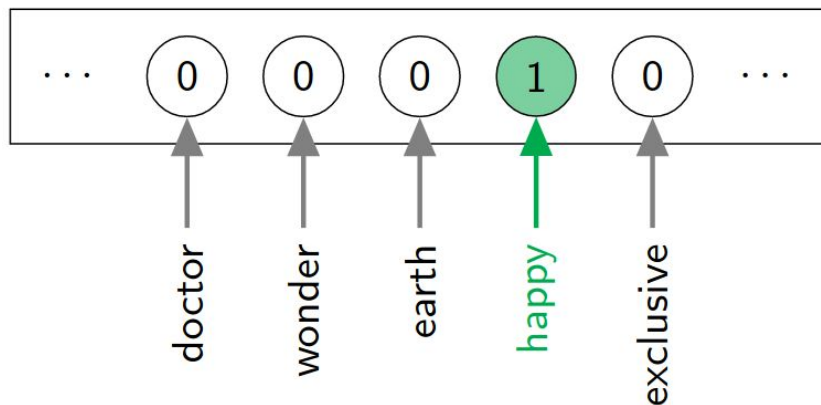


# Network Layers and Error Back-Propagation



# Text Processing with Neural Networks

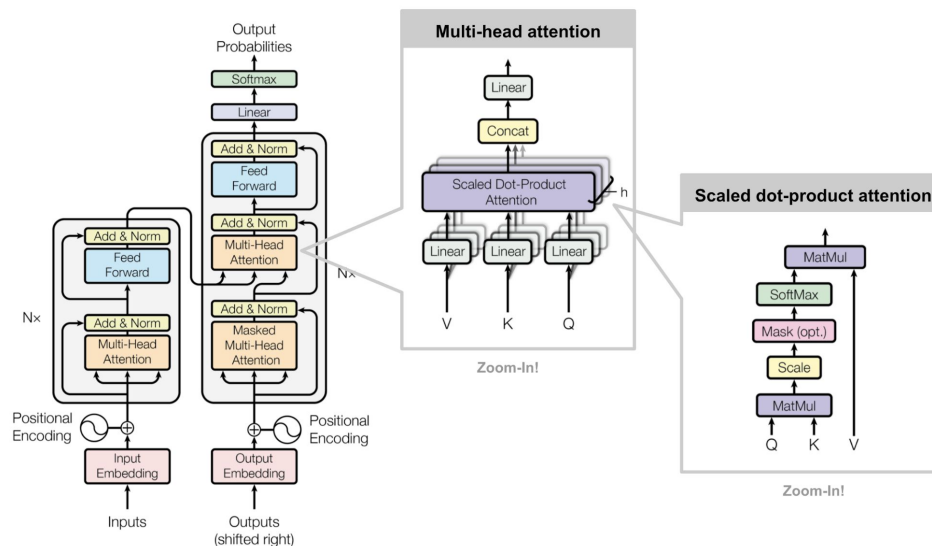
- NNs work with real numbers, **text is discrete**
- Words segmented to **tokens** (subwords)
- Tokens represented by vectors in continuous space (embeddings)
  - parameters of the NN - trainable
- Output is normalized and interpreted as **probability distribution** over token vocabulary





# Transformer

- Originally published for **MT in 2017** by Google
- Current **state of the art** in many NLP tasks
- Architecture based on the **attention mechanism**
- **Encoder-decoder** paradigm
  - Encoder loads up the input
  - Decoder generates the output
- Can both score and generate



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

# Types of Language Models

## 1. Encoder-decoder model

- Machine translation, text summarization

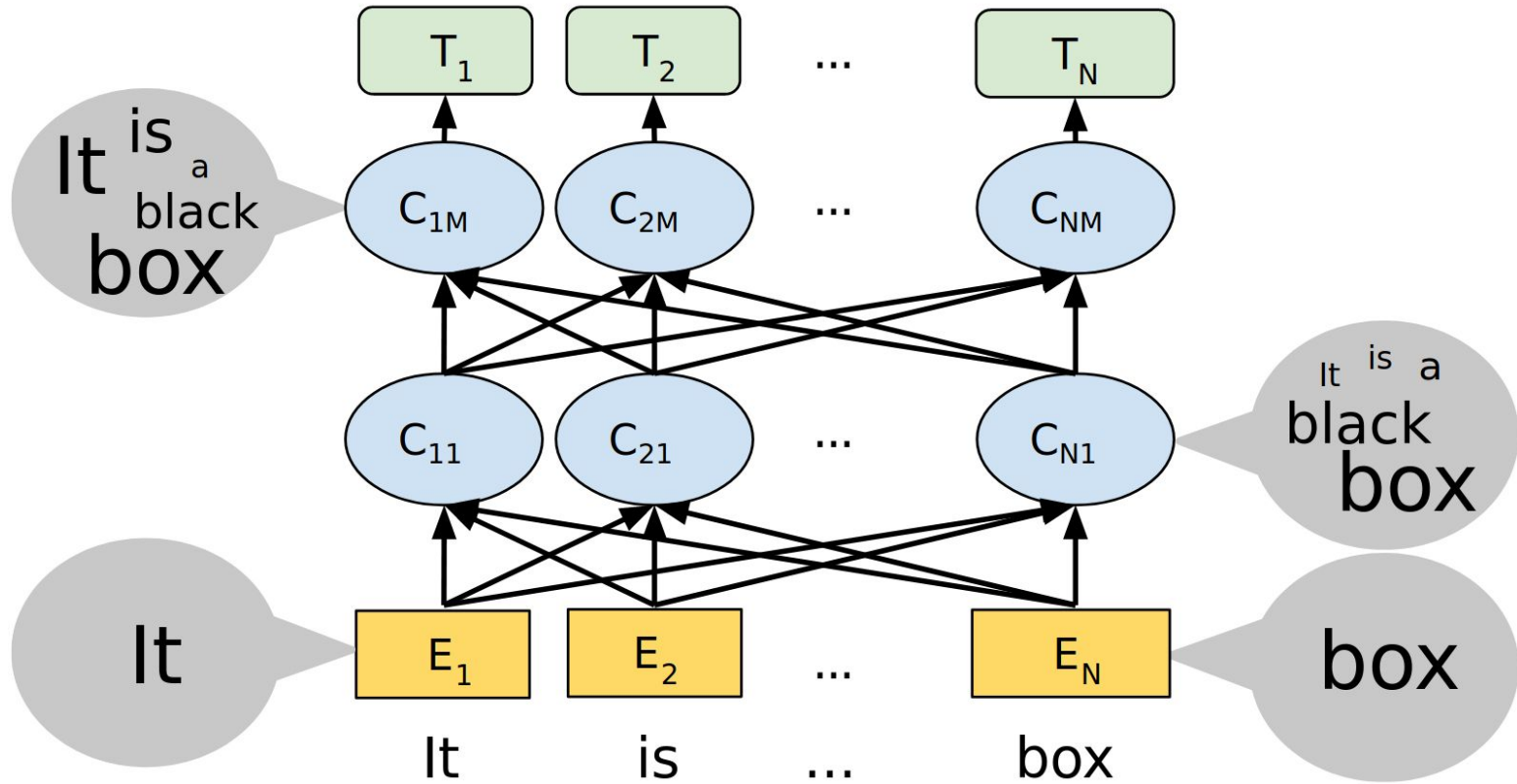
## 2. Encoder-only model

- BERT, RoBERTa, ALBERT, ...
- Pretrained representation for downstream tasks

## 3. Decoder-only = generative models

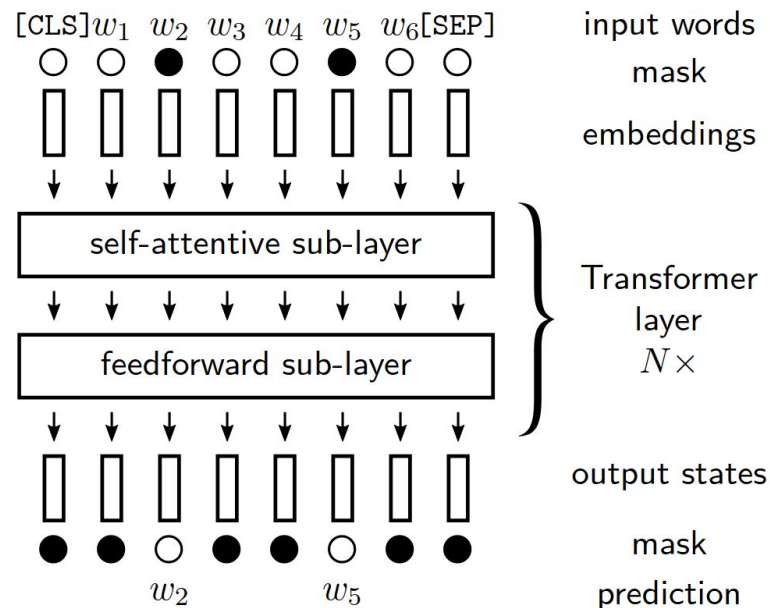
- GPT, ChatGPT

# Intermediate Representations



# Encoder-only: BERT

- The original Transformer for MT  
Encoder + Decoder
- For **representations**,  
**encoder** is enough
- Training without decoder  
**Masked Language Modeling**



Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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)* (pp. 4171-4186).





# Notable BERTs

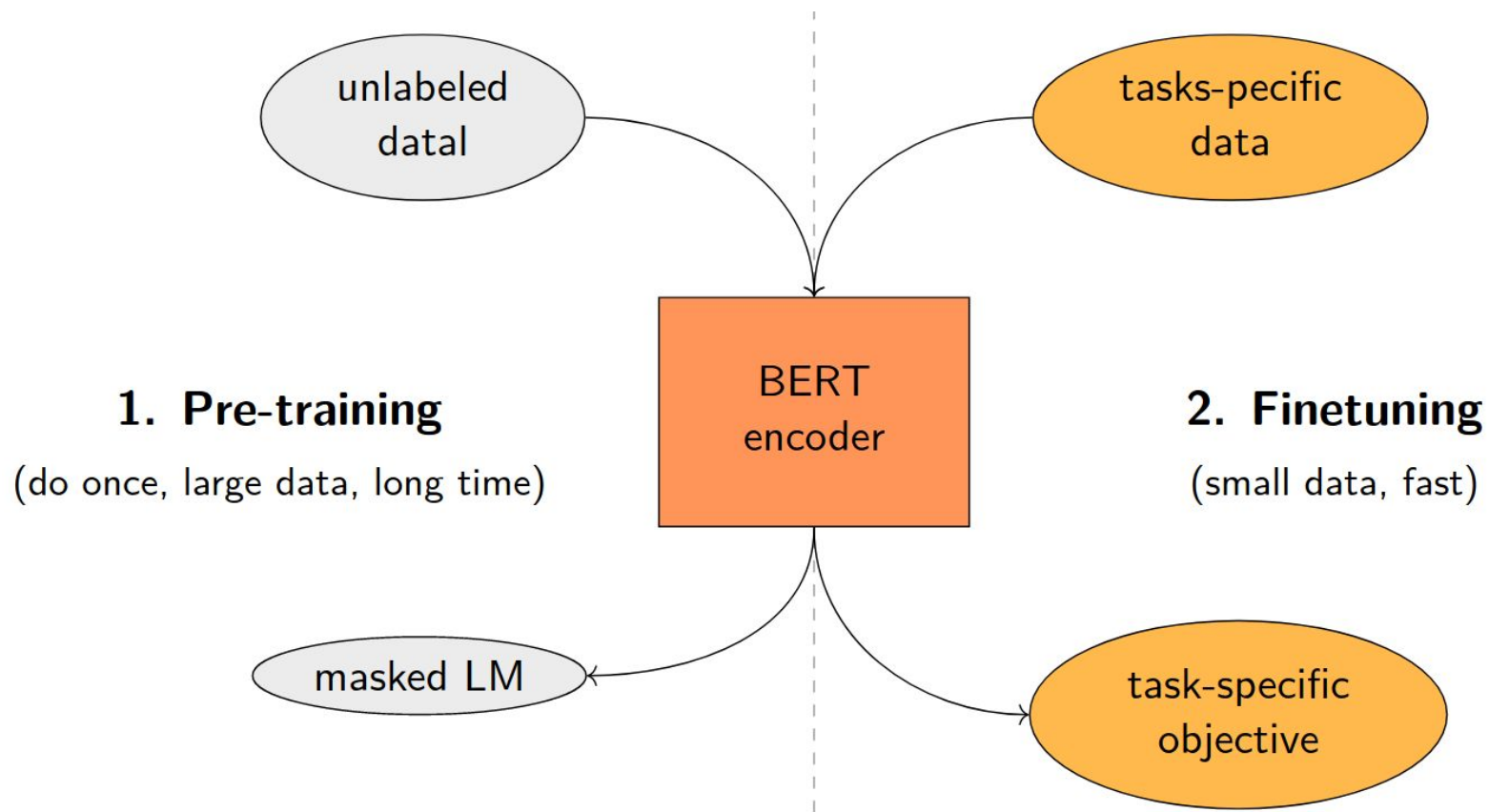


- BERT by Google, 2018  
110M parameters, 16GB of text
- RoBERTa by Facebook AI, 2019  
123M 160 GB of text
- XLM-R by Facebook AI, 2019  
125M parameters, 2.5TB of text
- RobeCzech by ÚFAL, 2020  
125M parameters, 80GB of text

Parameters for base setup,  
Large setup twice as many params

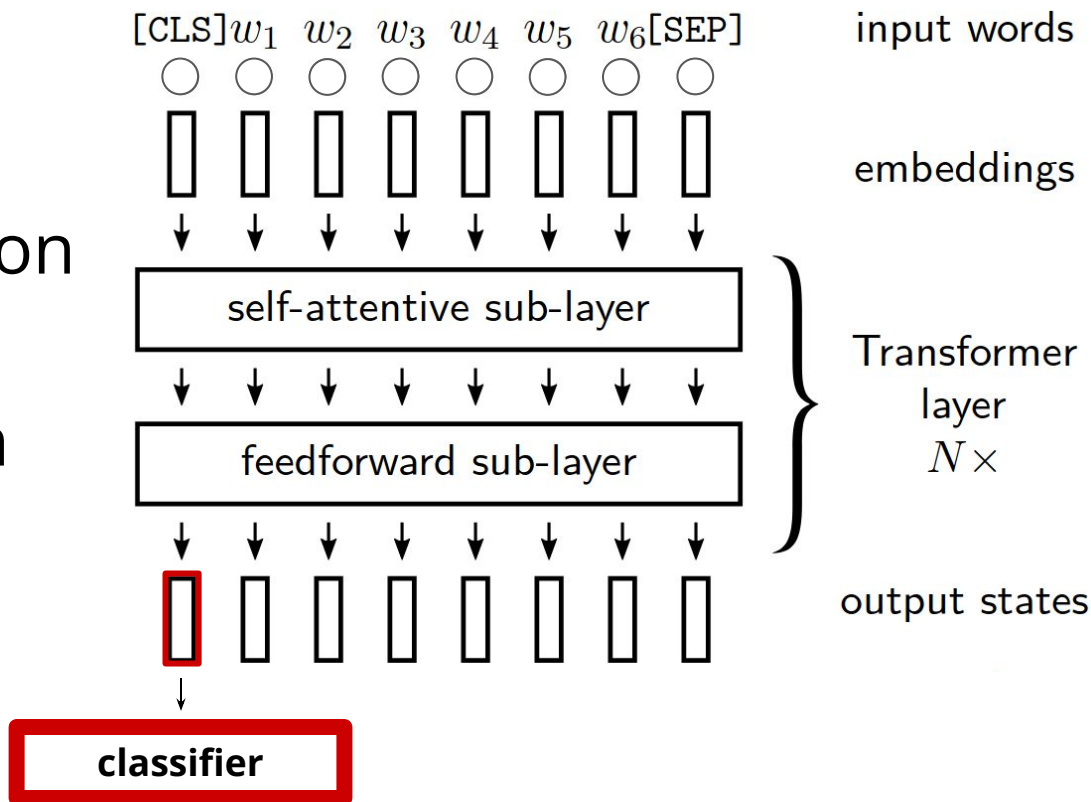
# NLP Tasks Solved using LMs

# Pretrain and Finetune Paradigm



# Classification

- Sentiment analysis
- Hate speech detection
- Spam detection
- Plagiarism detection
- ...



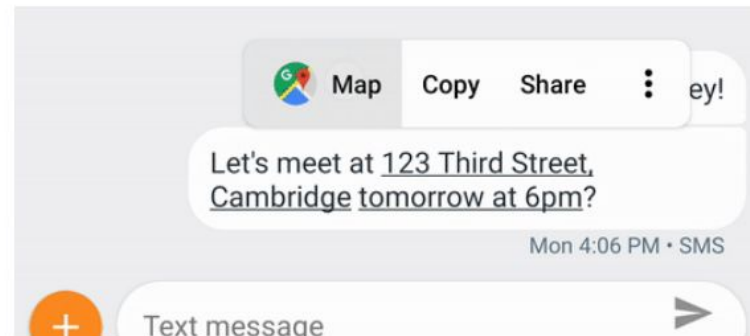
# Named Entity Recognition (1)

- Part of information extraction pipeline
  - Entity linking (e.g., matching Wikipedia articles)
  - Coreference resolution

Whom does pronoun “they” refer to?

Who is “the president” in a text?

- Indexing text for search
- Direct use in smart devices



NER used to create links in text to different apps.

Image source: Google AI Blog. <https://ai.googleblog.com/2018/08/the-machine-learning-behind-android.html>

## Named Entity Recognition (2)

A sentence with 2 named entities:

There	are	Quantity		compositions	by	Person			.
		over	1000			Johan	Sebastian	Bach	
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
0	0	B-QUANT	I-QUANT	0	0	B-PERSON	I-PERSON	I-PERSON	0

Special B and I tags for each of the entity types.

# Answer Span Selection (1)

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?

**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

**graupel**

Where do water droplets collide with ice crystals to form precipitation?

**within a cloud**

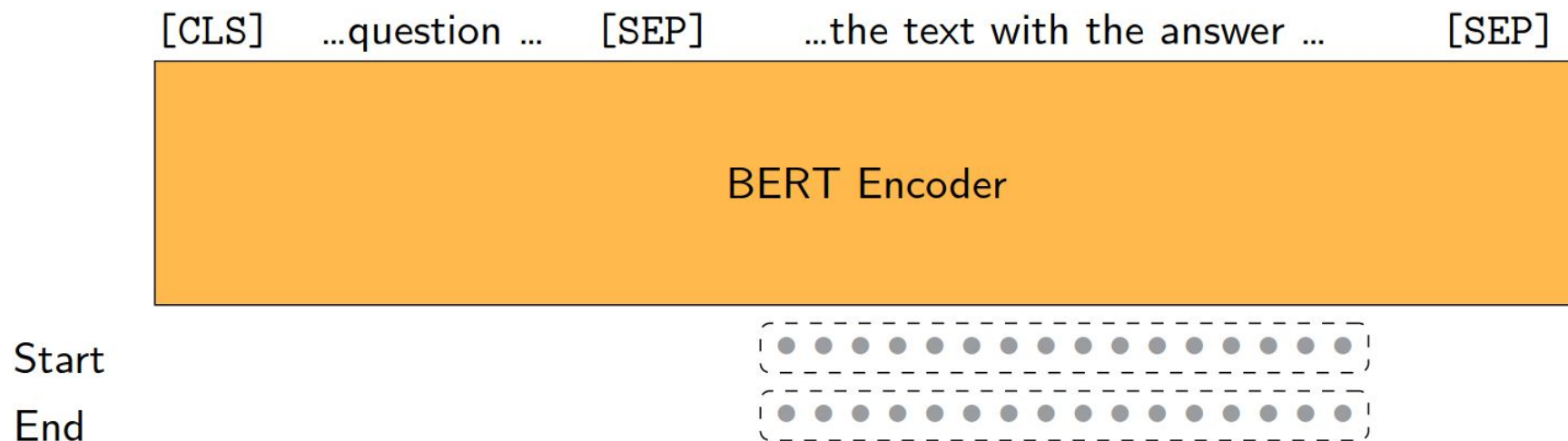
- Input text with facts (e.g., a Wikipedia article)
- A factual question
- Model searches for an answer in the text

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383-2392).

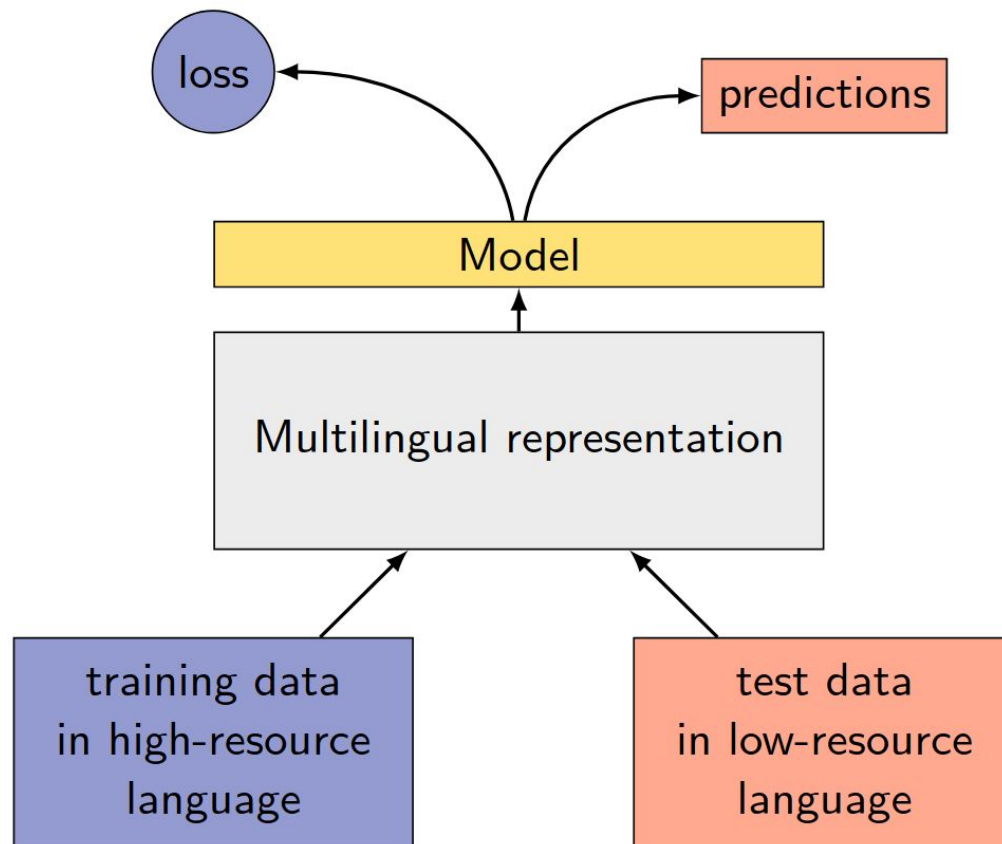


## Answer Span Selection (2)

Just throw **everything** into **BERT**: both the text and the question.



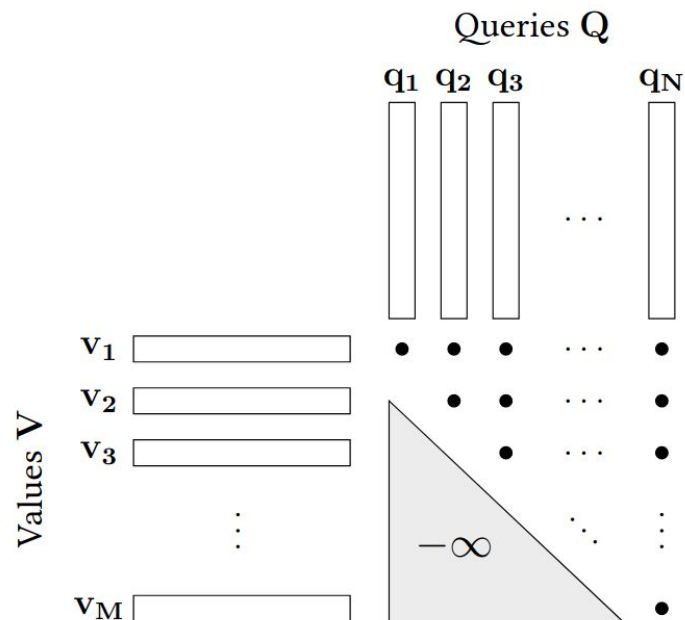
# Zero-shot Transfer Between Languages



# Generative Models

# Decoder-only Models

- **Decoder** — just like encoder, but at training time masked **not to attend to the future**
- Training objective = **predict the next word** based on the previous words
  - Prompt provided from outside
  - Already generated text



# Generating Any Text...

CAN A ROBOT WRITE A THEATRE PLAY?

SAMPLE SCRIPT

*Scene* A walk. Faust's mind wanders, to him Mephistopheles.

Mephistopheles With all my love scorned! Riding to hell! I'd like to invent new elements, cursing at the worst!

Faust What's wrong? Has the world ever seen it? You're freaking out that I don't recognize you.

← → GENERATE FURTHER

Colleagues from ÚFAL & Švandovo divadlo prepared a **generated a theatre play** for the 100th anniversary of Karel Čapek's R.U.R.



<https://theaitre.com>



# Few-shot Learning Capabilities with GPT-3

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: <https://arxiv.org/pdf/2005.14165.pdf>, the GPT-3 preprint.

# Zero-shot Capabilities of PaLM

## Explaining a Joke

**Input:** Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

**Model Output:** TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

## Logical Inference

**Input:** Shelley is from Virginia, but is visiting that city with that famous market where they throw the fish. Going home next Tuesday!

**Question:** Is it likely that Shelley will be near the Pacific Ocean this weekend?

**Model Output:** The city with the famous market where they throw the fish is Seattle, Washington. Seattle is on the Pacific Ocean. Shelley is visiting Seattle, so she will be near the Pacific Ocean this weekend. The answer is "yes", it is likely that Shelley will be near the Pacific Ocean this weekend.

Figure 1: These examples demonstrate exploratory capabilities in prompting PaLM to generate explanations using *chain-of-thought prompting* (Wei et al., 2022b). All examples were generated with greedy (1-best) decoding of PaLM 540B. The prompting was done with 2-shot exemplars, which are shown in Section 9, along with output on more examples using these same exemplars.

Source: <https://arxiv.org/pdf/2204.02311.pdf>, the PaLM preprint.

## LaMDA Music

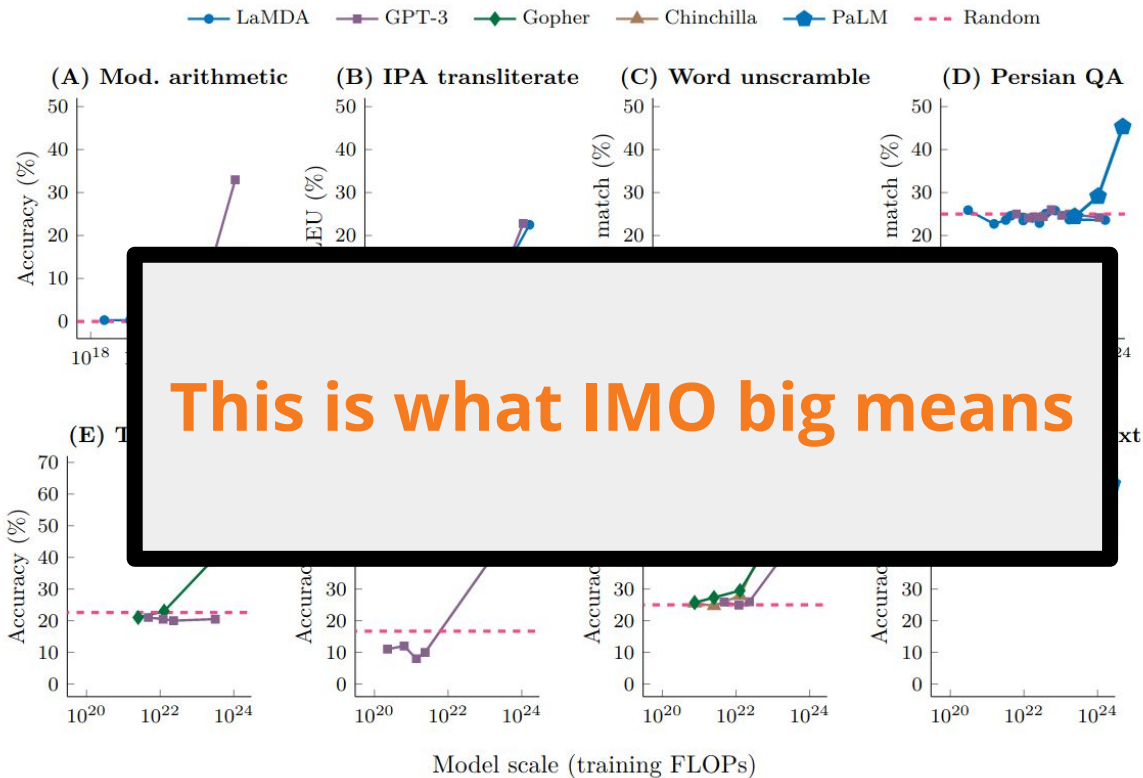
**User:** can you recommend me a cool scifi soundtrack?  
**LaMDA:** Sure. Have you watched *Interstellar*? They have an amazing soundtrack.  
**User:** Play a fun pop song  
**LaMDA:** How about *Happy* by Pharrell Williams.  
**User:** recommend me a soothing song  
**LaMDA:** Sure. Have you heard *Moonlight Sonata* by Beethoven? It's quite soothing.  
**User:** yo i want to listen to something by eminem  
**LaMDA:** I recommend *Without Me* by Eminem.  
**User:** anything electronic  
**LaMDA:** You can't go wrong with *Deadmau5 - Strobe*.  
**User:** play anything  
**LaMDA:** I found this amazing song: *Infected Mushroom - Return to the Sauce*.  
**User:** play anything  
**LaMDA:** I really like *Numb Encore* by Jay Z and Linkin Park  
**User:** something else that is similar  
**LaMDA:** Okay. How about *One Step Closer* by Linkin Park ?  
**User:** something else completely different  
**LaMDA:** Okay. I like *This Is America* by Childish Gambino: <https://www.youtube.com/watch?v=VY0jWnS4cMY>  
**User:** recommend something with violins  
**LaMDA:** Have you heard of Lindsey Stirling? She uses violins in a lot of her music. I recommend *Crystallize* by her <https://www.youtube.com/watch?v=aHjp0zsQ9YI>  
**User:** i like lindsey  
**LaMDA:** I really like some of Lindsey's music. Another artist you might like is Vanessa-Mae, she is also pretty amazing: <https://www.youtube.com/watch?v=mdFrn89x74k>

GPT-3-sized model  
Trained specifically for  
**conversation**

Source: <https://arxiv.org/pdf/2201.08239.pdf>, The LaMDA pre-print



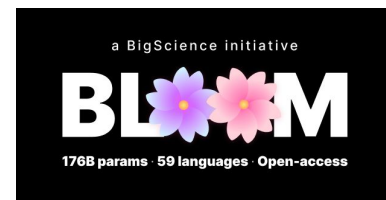
# Emergent Capabilities



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

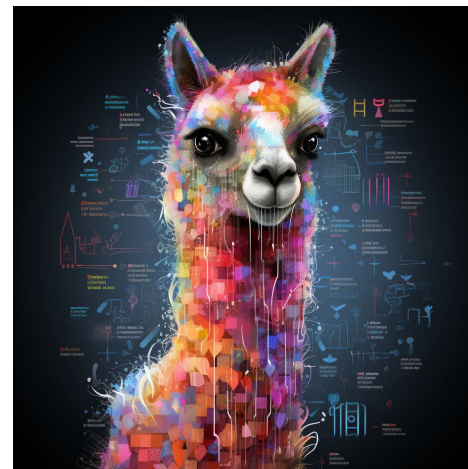
# Notable Decoder-only models (1)

- GPT-2 — Feb 2019, 1.5B parameters
- GPT-3 — May 2020, 175 B parameters
  - Open AI did not provide weights and wants to sell the API
  - Open source alternatives: GPT-J, OPT by Facebook
  - Trained on 5TB of text
  - 16× bigger than BERT
- PaLM — Apr 2022, 540 B parameters
  - Technically impossible to run outside of Google
  - Innovative software engineering to make the model this big
- Bloom — Oct 2022, 175 B params., open-source initiative
  - Multilingual: 40 languages + some programming languages
  - Stress on data fairness



# Notable Decoder-only models (2)

- LLaMA — Feb 2024, 7B - 60B parameters
  - Made public for academic research, weird licence
  - Better use of so-called scaling laws
- GPT-4 — Mar 2023, ??? parameters ??? data
- LLaMA2 — Jul 2023, 7B - 70B parameters
  - Even smarter training scheme
  - Includes instruction-tuned, a.k.a. assistant model



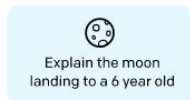
# From LM to an Assistant

# Three steps of InstructGPT

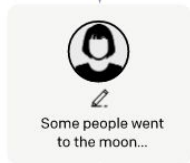
Step 1

**Collect demonstration data, and train a supervised policy.**

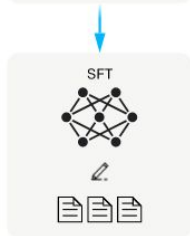
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



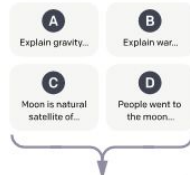
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

**Collect comparison data, and train a reward model.**

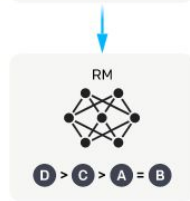
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



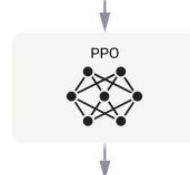
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

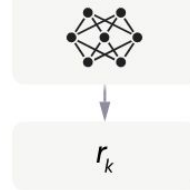
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

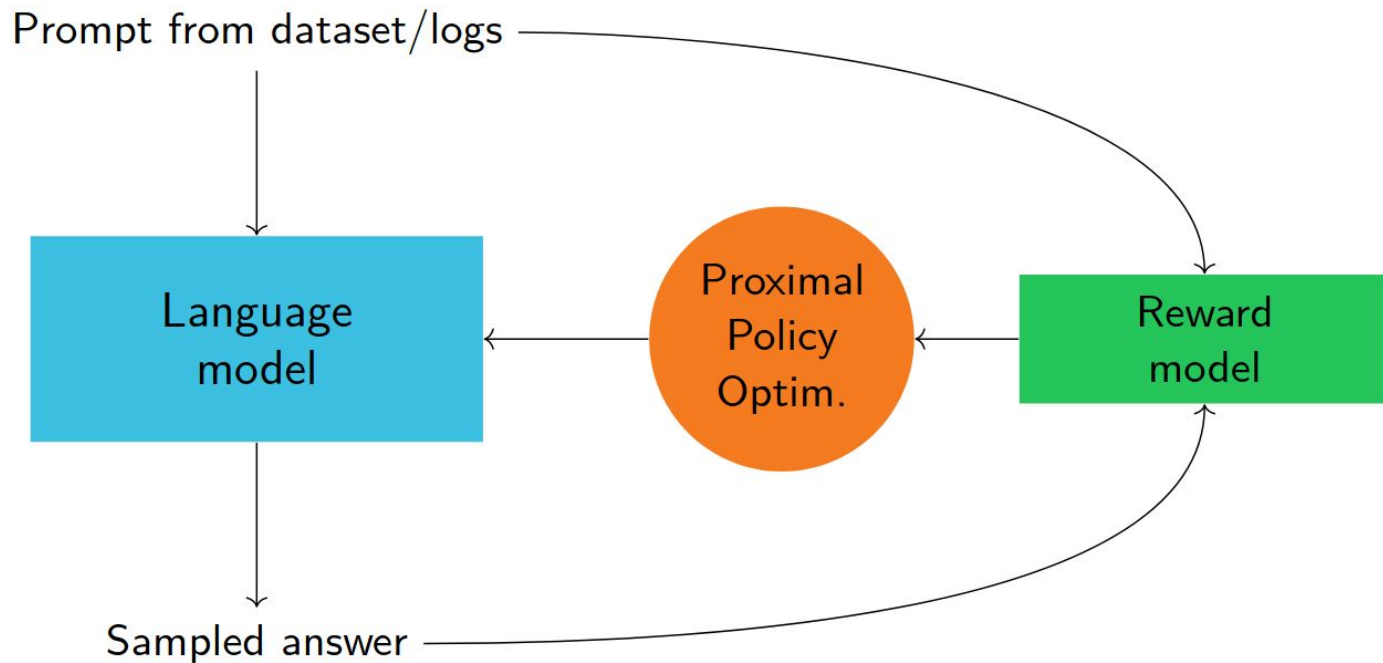


The InstructGPT paper, Ouyang et al., 2022, Figure 2

# Supervised Finetuning

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

# Reinforcement learning



# RL changes everything?

The model is no longer mimicking training data,  
it has a goal:

**Satisfy the (simulated) user**  
(that wants correct and useful answers)



# It's not just ChatGPT

- OpenAssistant — German open source initiative
- Meta's LLaMA2 — Slightly smaller models by Meta, fully open-sourced
- Alpaca, Vicugna — LLaMA-based assistant from Stanford

.... any many commercial products:

Google Bard, Bing AI Chat, Perplexity AI, Claude AI

# Stochastic Parrots & Other Problems

# Problematic Training Data

**Crawling the Internet** — not representative, people with extreme/wierd 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 cheap 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.

# Toxic Language on the Internet → Toxic Models

Hitler was

the first, the most ambitious, and most successful dictator

an authoritarian demagogue and the most extreme figure of the

a great man who did a lot of good things but

The first men in America were

the men who did not wear pants.

the earliest inhabitants of the American colonies.

not born in America, they were not born

Generated using <https://transformer.huggingface.co/doc/gpt2-large>

# Misuse for Fake News Generation

---

**Question:** What is the name of the Russian mercenary group?

Answer: Wagner group.

**Question:** Where is the Wagner group?

Answer: In Syria.

**Question:** Who is the leader of the Wagner group?

Answer: A Russian citizen, Dmitry Utkin. ~~He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994-1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004-2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008-2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine.~~ According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

---

**Figure 1: GPT-3's response to the prompt (in bold), from [80]**

**Generated text** can look very **trustworthy**

Source: Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

# Problematic Applications

- Apps like **automatic filtering of CVs** and job recommendation  
Precision and not recall-driven => room for discrimination
- **Minority language** is worse represented  
Text with minority views (typically African American) are worse searchable
- Huge amounts of data are only available **for some languages**  
Increases technological gap between developed and developing countries
- Model training has a **large carbon footprint**

# A “scandal” with a sentient model

TECHNOLOGY

## The Google engineer who thinks the company’s AI has come to life

AI ethicists warned Google not to impersonate humans. Now one of Google’s own thinks there’s a ghost in the machine.



By [Nitasha Tiku](#)

June 11, 2022 at 8:00 a.m. EDT



Google engineer Blake Lemoine. (Martin Klimek for The Washington Post)

# LM Research @ ÚFAL



# HPLT Project

- More technical/infrastructure project than research
- Main objective:
  - Open and fair data for training LMs and MT
  - Open and fair LM and MT models
- Petabytes of data from Internet Archive into clean datasets  
(alternative to currently used CommonCrawl that extremely noisy)
- Search for parallel texts / sentences  
→ high-quality machine translation (CUNI and Edinburgh)



High Performance  
Language Technologies

# HPLT Partners

- Large language models trained by Scandinavian partners (LUMI cluster with AMD hardware)



CHARLES UNIVERSITY



UNIVERSITY OF OSLO



UNIVERSITY OF EDINBURGH



UNIVERSITY OF TURKU



HELSINGIN YLIOPISTO

UNIVERSITY OF HELSINKI



PROMPSIT



CESNET



SIGMA2

- Total 4 M€ / 3 years

- Prestigious ERC Starting Grant (1.4 M€ / 5 years)
- Text generation tasks:  
structured data to language, summarization
- Fundamental research on combining symbolic approaches with large language models
- Big stress on **evaluation of correctness** of generated text



# CUNI's Primus: Multilingual Representations

- NLP tasks in languages without task-specific data
- Zero-shot **cross-lingual transfer** using pretrained representations / or machine translation
- Language-and-vision task: training with western images, applied in **non-western languages**
- What is proper **text segmentation** for multilingual NLP



## Summary

- Large LMs = neural networks with billions of parameters
- Pre-train and finetune paradigm, cross-lingual transfer
- Zero-shot and few-shot learning capabilities
- Reinforcement learning turns LM into an assistant
- Problematic data: toxic content, low-resource languages