Advanced NLP -

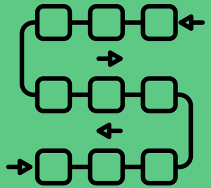
Session 5: LLMs

Prof. Dr. Richard Sieg
TH Köln IWS - WS 25/26

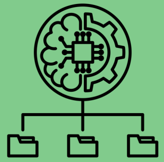
Lecture	Date	Topic
1	03.12.2025	Introduction & NLP Recap
2	04.12.2025	RNNs and LSTMs
3	10.12.2025	Attentions & Transformers
4	11.12.2025	Transformer Based Models
5	17.12.2025	Hackathon / Check-In
6	18.12.2025	LLM Architecture
7	07.01.2026	LLM Engineering
8	08.01.2026	Hackathon / Check-In
9	14.01.2026	LLM Shortcomings
10	15.01.2026	Final Presentations

Agenda

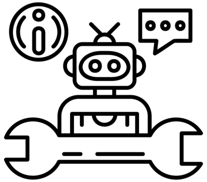
The 3 Ingredients of LLMs



Process long sequences and context



Efficient training on huge datasets



Follow (human) instructions

How do we get from this...

Write a poem about Bruce Lee.

I love this guy. His poems are very original and have a very vivid visual language to them and the way he talks about this country.

GPT2

...to this

Write a poem about Bruce Lee.

And when he moved, we saw
not just a man fighting
but a principle made flesh:

Be water, my friend.

Sonnet 4.5

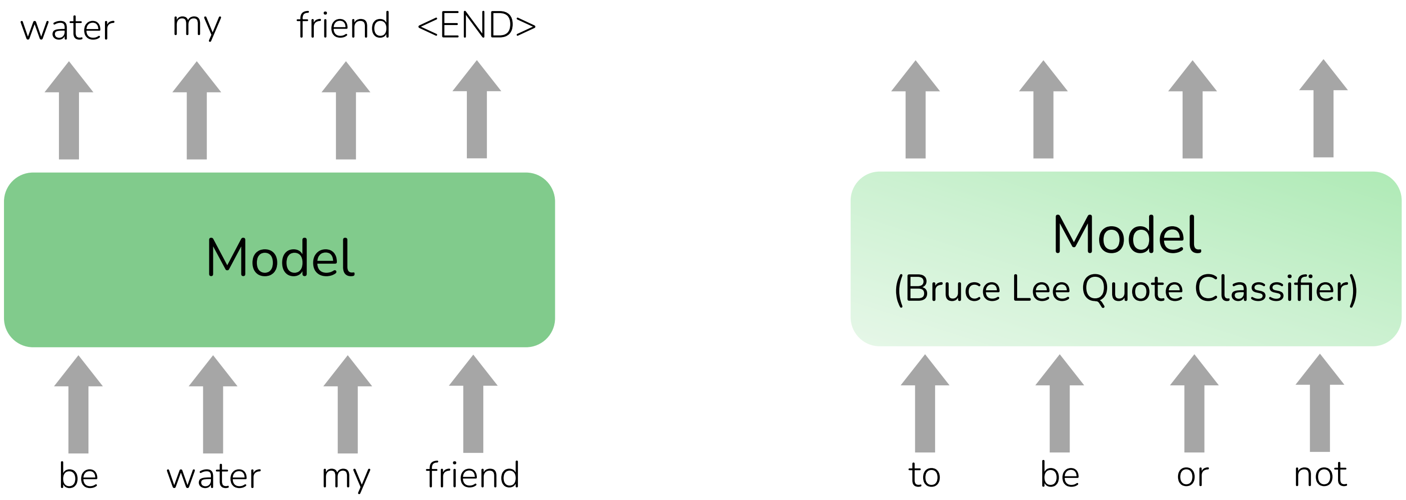
01.

Instruction Fine-Tuning

Pre-Training Paradigm (last lecture)

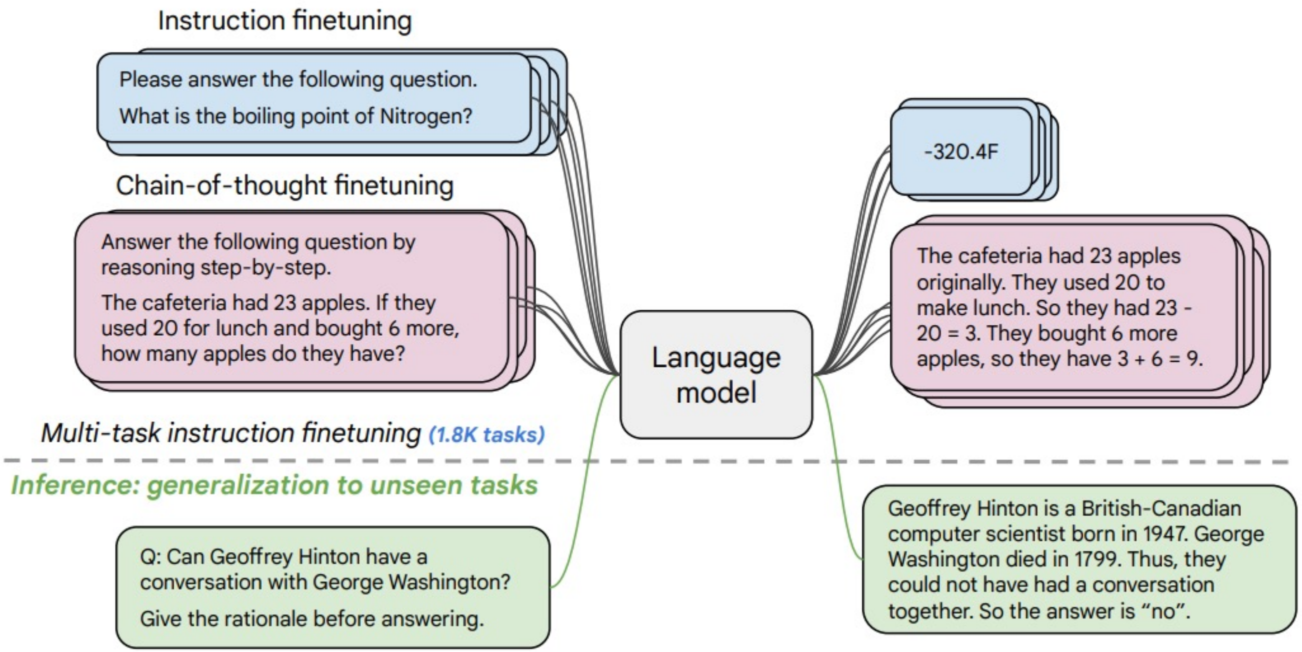
- Nowadays, we take a pre-trained model for our language (or a multilingual model) and use this as initial parameters to train our downstream task

Step1: Pre-train model on lots of text   Step2: Fine-tune on ***many tasks***



Instruction Fine-Tuning

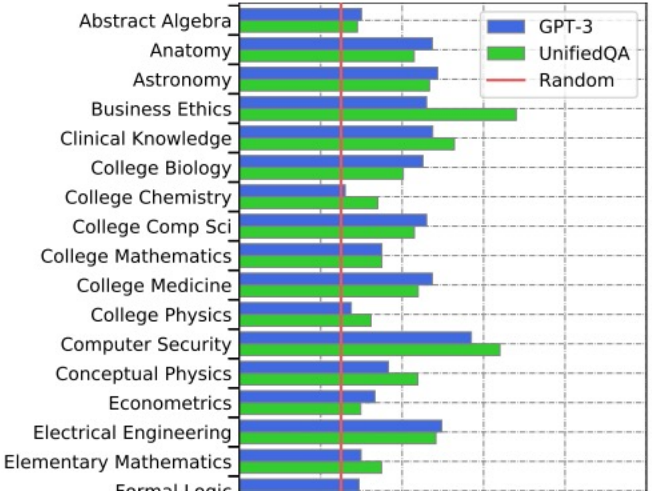
- Fine-tune LM on instruction → output pairs and evaluate on unseen tasks



[Scaling Instruction-Finetuned Language Models (Flan-T5) [Chung et al 2022]]

Multitask Benchmarks - MMLU

- Massive Multitask Language Understanding
- Contains 57 diverse knowledge tasks
- Was/Is the “go to” LLM benchmark (up until shortly)
- Benchmarks can’t keep up with LLM development



Why isn’t there a planet where the asteroid belt is located?
(A) A planet once formed here but it was broken apart by a catastrophic collision.
(B) There was not enough material in this part of the solar nebula to form a planet.
(C) There was too much rocky material to form a terrestrial planet but not enough gaseous material to form a jovian planet.
(D) Resonance with Jupiter prevented material from collecting together to form a planet.

Figure 16: An Astronomy example.

If each of the following meals provides the same number of calories, which meal requires the most land to produce the food?
(A) Red beans and rice
(B) Steak and a baked potato
(C) Corn tortilla and refried beans
(D) Lentil soup and brown bread

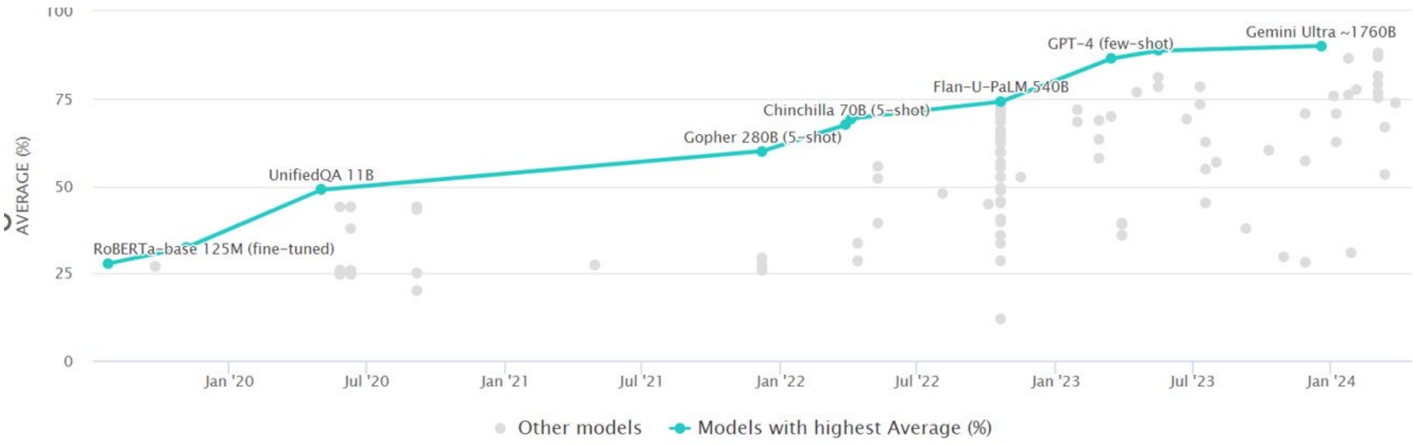
Measuring Massive Multitask Language Understanding [Hendrycks et al, 2020]

MMLU progress

GPT3 was SOTA

Model	Humanities	Social Science	STEM	Other	Average
Random Baseline	25.0	25.0	25.0	25.0	25.0
RoBERTa	27.9	28.8	27.0	27.7	27.9
ALBERT	27.2	25.7	27.7	27.9	27.1
GPT-2	32.8	33.3	30.2	33.1	32.4
UnifiedQA	45.6	56.6	40.2	54.6	48.9
GPT-3 Small (few-shot)	24.4	30.9	26.0	24.1	25.9
GPT-3 Medium (few-shot)	26.1	21.6	25.6	25.5	24.9
GPT-3 Large (few-shot)	27.1	25.6	24.3	26.5	26.0
GPT-3 X-Large (few-shot)	40.8	50.4	36.7	48.8	43.9

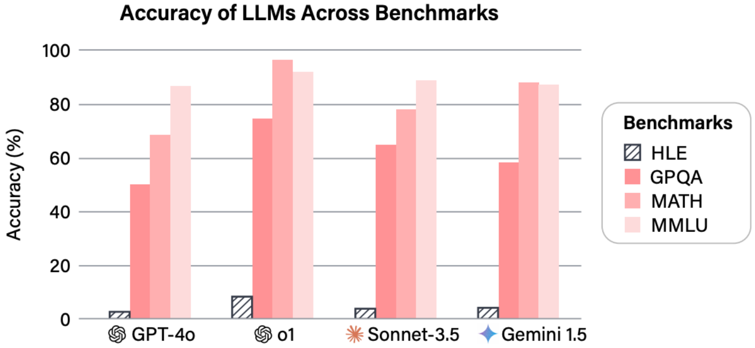
[LLM Stats Leaderboard]

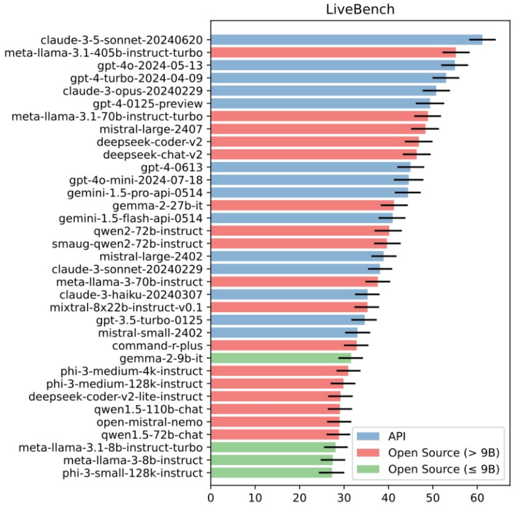


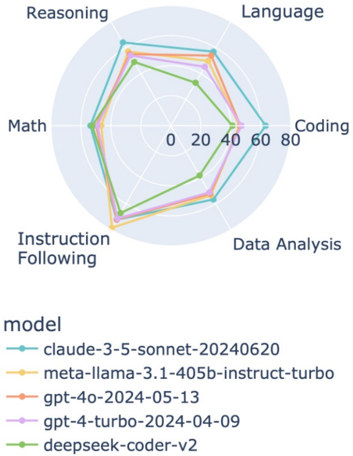
[Measuring Massive Multitask Language Understanding [Hendrycks et al, 2020]]

Some Other Benchmarks

- BIG-Bench (Google, 2022) contains 204 tasks
- Humanity's Last Exam (AI Safety, 2025)
- LiveBench (2024) gets updated monthly, to limit potential contamination into the training data of LLMs







Instruction Fine-Tuning Performance Gains

Params	Model	Norm. avg.	MMLU Direct	MMLU CoT	BBH Direct	BBH CoT	TyDiQA Direct	MGSM CoT
80M	T5-Small	-9.2	26.7	5.6	27.0	7.2	0.0	0.4
	Flan-T5-Small	-3.1 **(+6.1)**	28.7	12.1	29.1	19.2	1.1	0.2
250M	T5-Base	-5.1	25.7	14.5	27.8	14.6	0.0	0.5
	Flan-T5-Base	6.5 **(+11.6)**	35.9	33.7	31.3	27.9	4.1	0.4
780M	T5-Large	-5.0	25.1	15.0	27.7	16.1	0.0	0.3
	Flan-T5-Large	13.8 **(+18.8)**	45.1	40.5	37.5	31.5	12.3	0.7

Impact of Instruction Fine-Tuning increases with model size

-	Random	25.0
-	Average human rater	34.5
May 2020	GPT-3 5-shot	43.9
Mar. 2022	Chinchilla 5-shot	67.6
Apr. 2022	PaLM 5-shot	69.3
Oct. 2022	**Flan-PaLM 5-shot**	**72.2**
	Flan-PaLM 5-shot: CoT + SC	**75.2**
-	Average human expert	89.8

MMLU Scores

Model input (Snarks)

Q: Which statement is sarcastic?
Options:
(A) Wow thats crazy, I've never had a takeout meal that used hardwood containers before!
(B) Wow thats crazy, I've never had a takeout meal that used plastic containers before!
A: Let's think step by step.

PaLM 540B output

What is the speaker trying to say?
The speaker is trying to say that he has never had a takeout meal that used plastic containers before.
What is the definition of sarcasm?
the use of irony to mock or convey contempt.
What is the definition of irony?
the expression of one's meaning by using language that normally signifies the opposite [...]
(never stops generating)

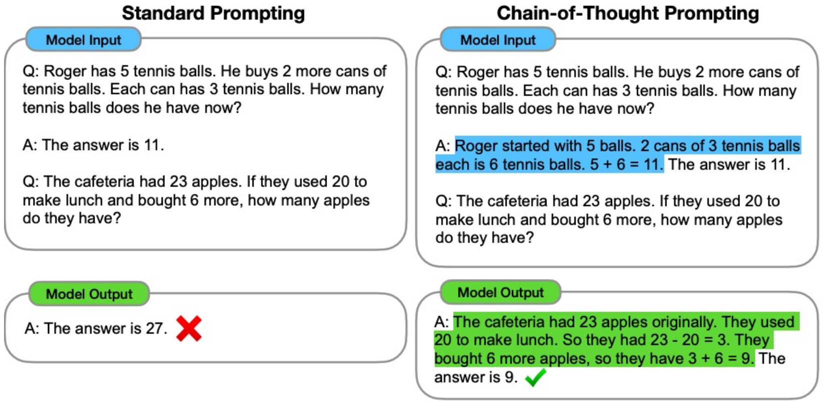
Flan-PaLM 540B output

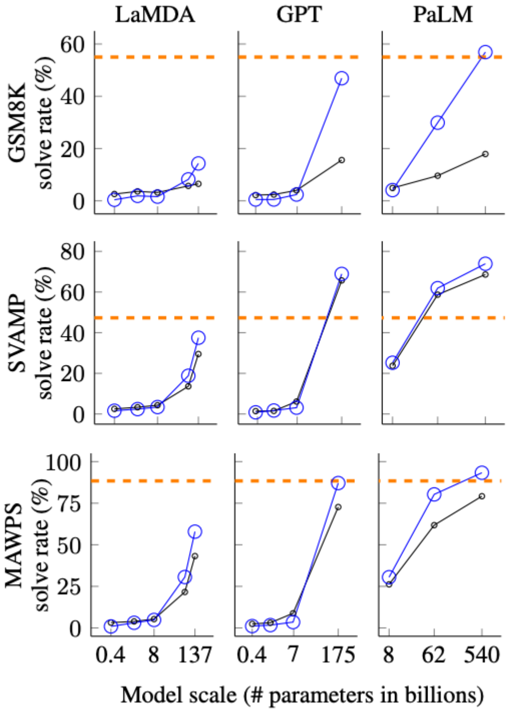
Plastic containers are very common for takeout meals. So, the answer is (B).

Scaling Instruction-Finetuned Language Models (Flan-T5) [Chung et al 2022]

Chain of Thought (CoT) Prompting

- Idea: Force the model to solve a problem step by step
- One simple but effective approach: Add “Let’s think step by step” in the beginning of the model output
- Nowadays we have reasoning models (GPT o1 familiy, DeepSeek, Gemini Pro etc)





Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [Wei et al, 2022]

Instruction Fine-Tuning Issues

- **Issue 1:** Lot of effort to create ground truth data and datasets are limited in size
  - One could use an LM for creating artificial datasets (e.g. Alpaca trained on 52k examples generated with LlaMA)
- **Issue 2:** There often is no *gold truth* answer to a task (e.g. "Write a poem.")
- **Issue 3:** Fine-tuning a language model penalizes ambiguous tokens the same as incorrect ones ("Avatar is a fantasy movie" vs "Avatar is an adventure musical")

How can we teach a model to follow human preferences?

02.

Reinforcement Learning from Human Feedback

Optimizing or Human Preferences

- We need a function $R(s)$ that measures human preference of a sample, the higher the better

```
SAN FRANCISCO,
California (CNN) --
A magnitude 4.2
earthquake shook the
San Francisco
...
overturn unstable
objects.
```

```
An earthquake hit
San Francisco.
There was minor
property damage,
but no injuries.
```

s_1

$R(s_1) = 8.0$

```
The Bay Area has
good weather but is
prone to
earthquakes and
wildfires.
```

s_2

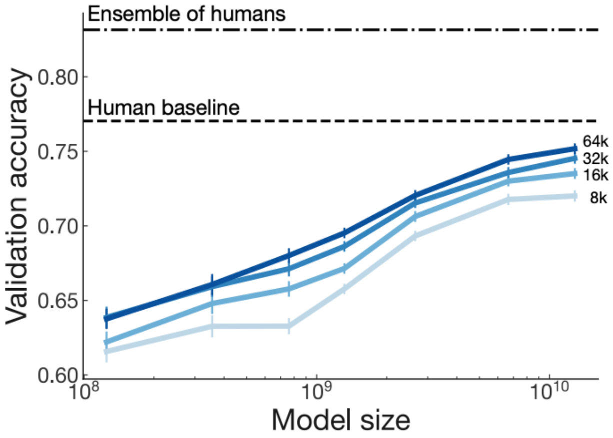
$R(s_2) = 1.2$

- Using ***Reinforcement Learning*** we can then optimize the parameters of our Language Model to *maximize* the expected reward of a sample set using ***gradient ascent*** (out of scope for this course)
- Optimization algorithm that OpenAI used: ***Proximal Policy Optimization*** [Schulman et al, OpenAI, 2017]

Deep reinforcement learning from human preferences [Christiano et al, Google, 2017]

Modeling Human Preferences – Finding $R(s)$

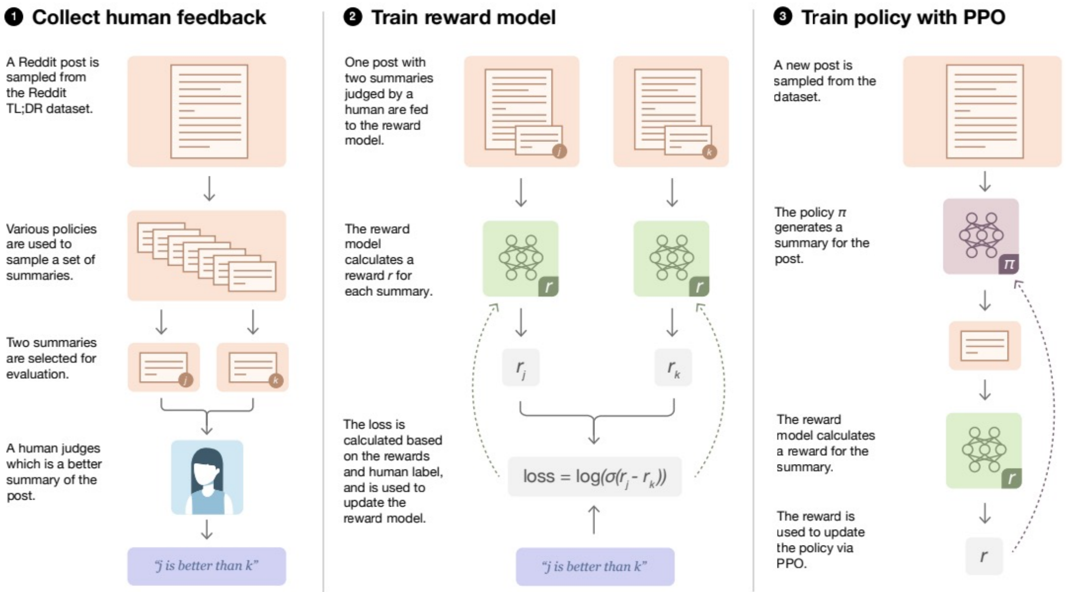
- **Issue 1:** Asking humans to label millions of samples is to expensive
- **Solution:** Train a Language Model $\mathrm{RM}_\Phi(s)$ to predict human preference
- **Issue 2:** Direct human ratings are noisy and miscalibrated
- **Solution:** Ask for pairwise comparisons and train $\mathrm{RM}_\Phi(s)$ to rank wining models higher

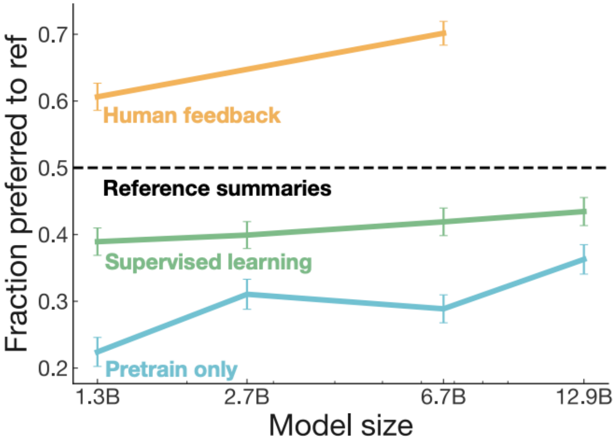


Learning to summarize from human feedback [Stiennon et al, OpenAI, 2020]

RLHF for Summaries

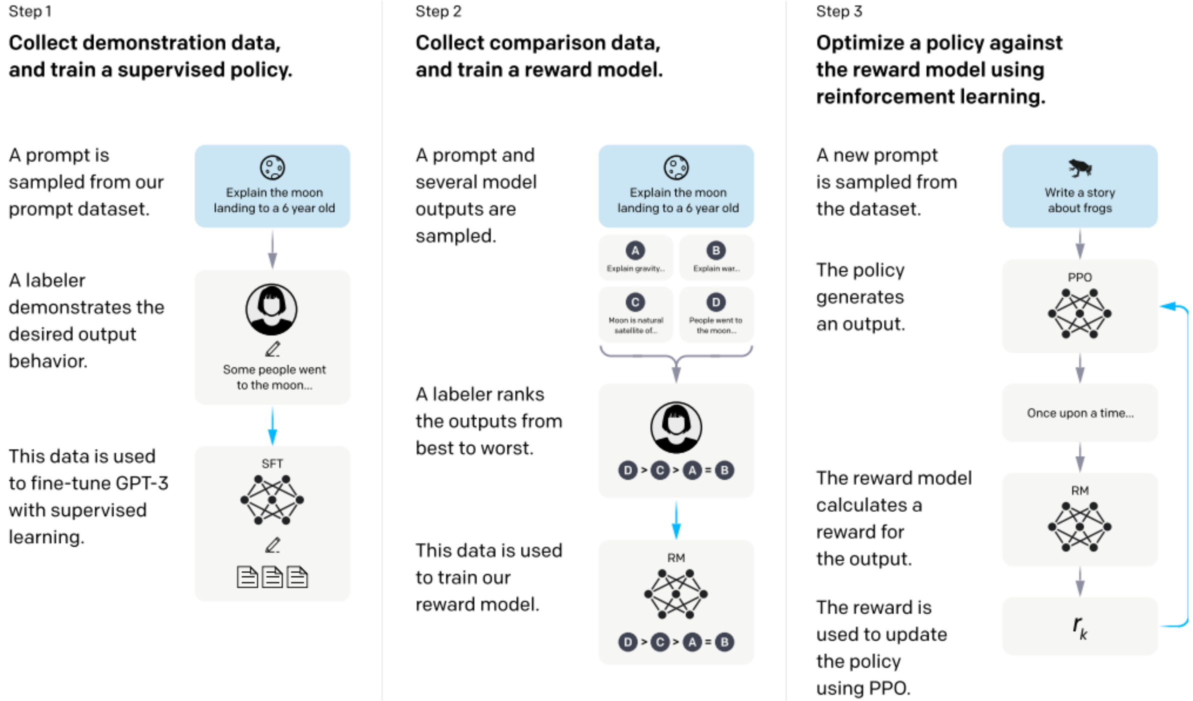
- We can now put everything together and optimize the parameters of our model using Reinforcement Learning (while not diverging too far away from the pretrained model – detail)





[Learning to summarize from human feedback [Stiennon et al, OpenAI, 2020]]

InstructGPT: Scaling RLHF up



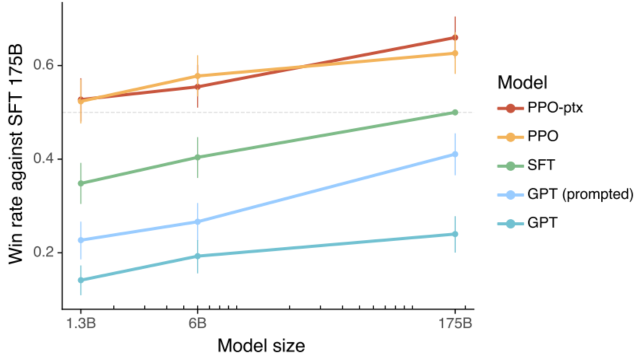


Table 6: Dataset sizes, in terms of number of prompts.

SFT Data			RM Data			PPO Data		
split	source	size	split	source	size	split	source	size
train	labeler	11,295	train	labeler	6,623	train	customer	31,144
train	customer	1,430	train	customer	26,584	valid	customer	16,185
valid	labeler	1,550	valid	labeler	3,488			
valid	customer	103	valid	customer	14,399			

[Training language models to follow instructions with human feedback [Ouyang et al, OpenAI, 2022]]

ChatGPT

- Same technique as InstructGPT with a focus on dialogue and using GPT 3.5

Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

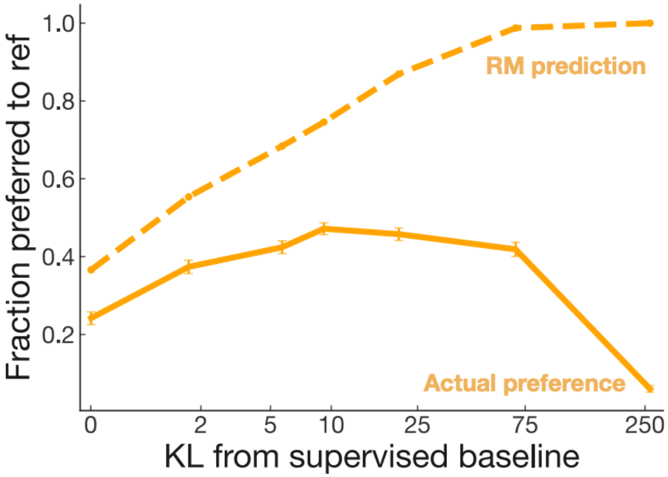
We have our first popular LLM

03.

Direct Preference Optimization

RLHF Issues

- “***Reward Hacking***” (common RL problem) i.e. overfitting to the reward
- Need to train and manage another model
- Human preferences are noisy
- Leads to *hallucinations* and *sycophancy* (answers seem to be helpful/truthful or agree too much with the user)
- Goodhart’s Law: “*When a measure becomes a target, it ceases to be a good measure*”



Direct Preference Optimization

- Clever people found out that there is a closed form solution to the Reinforcement Learning objective
- Using some math, we can come up with a loss function to fine-tune our LM

LM to fine-tune

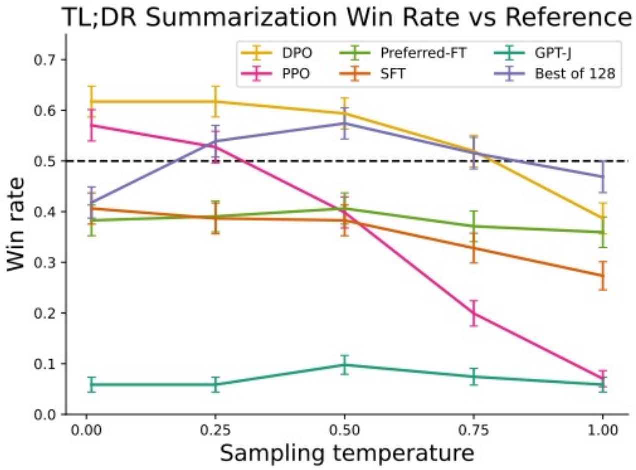
$$\mathcal{L}_{\mathcal{DPO}}(\pi_\theta; \pi_{\text{ref}}) =$$

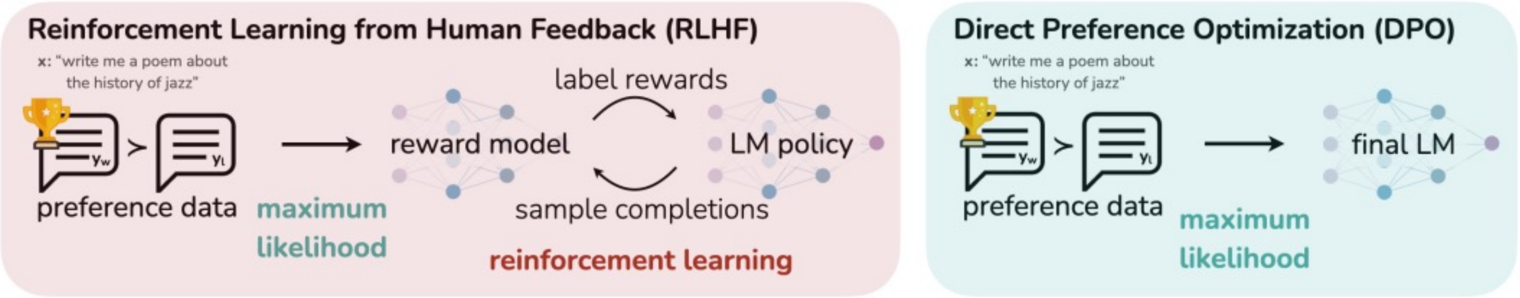
Frozen LM
(Pre-Trained & SFT)

Direct Preference Optimization: Your Language Model is Secretly a Reward Model [Rafailov et al, 2023]

Direct Preference Optimization

- DPO is now the go-to algorithm for open-source LLMs





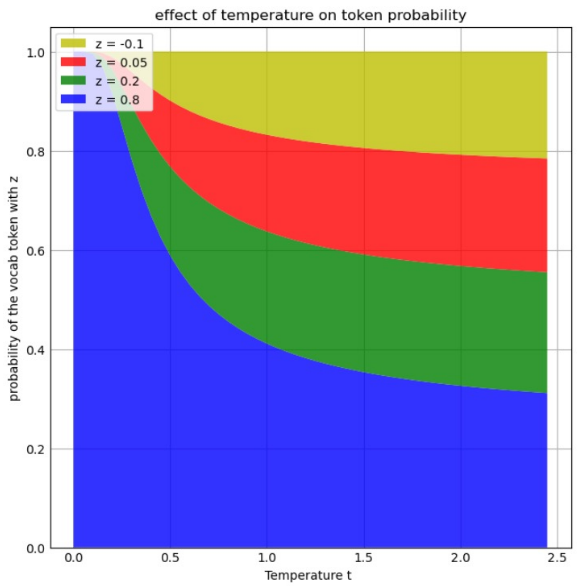
Direct Preference Optimization: Your Language Model is Secretly a Reward Model [Rafailov et al, 2023]

04.

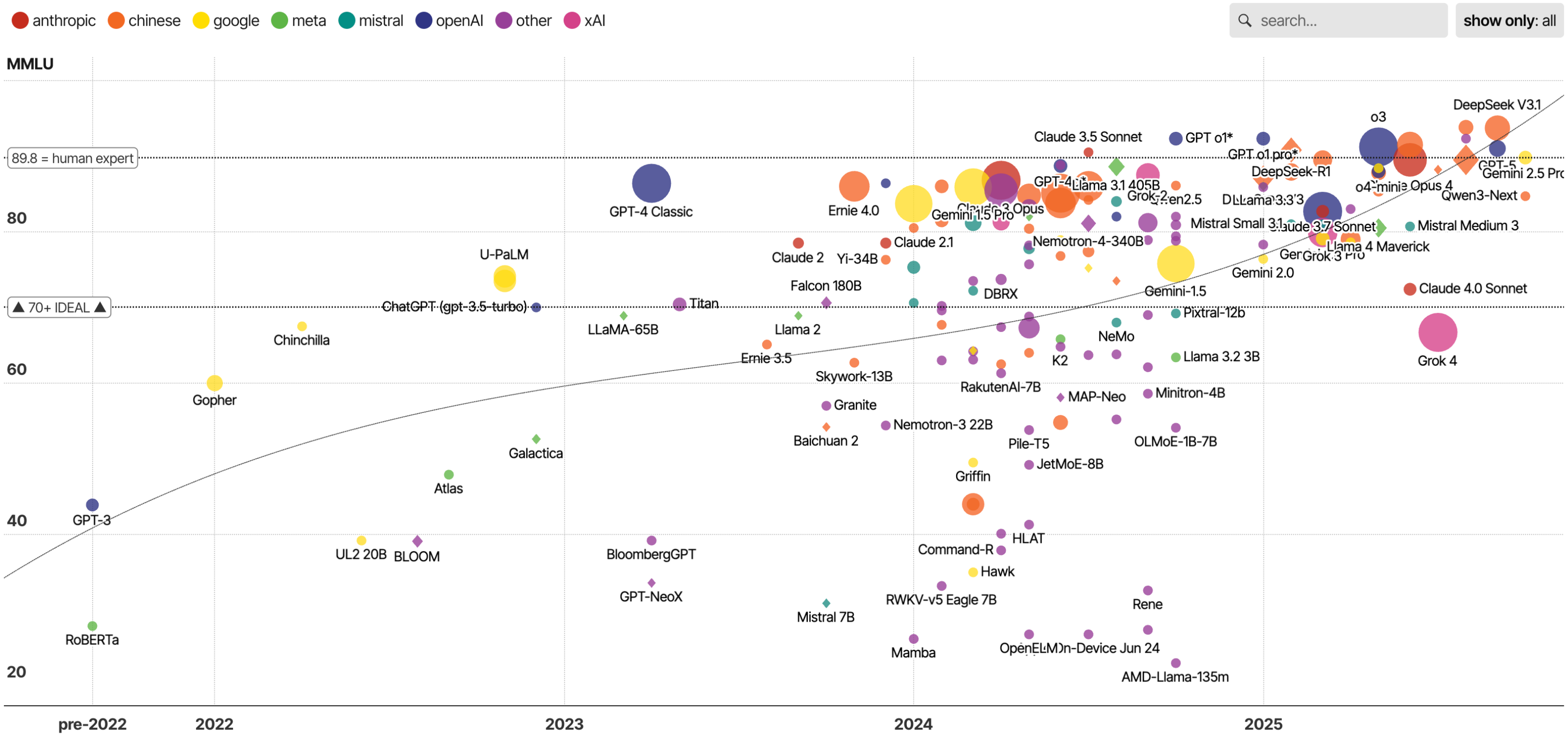
LLM Landscape

LLM Generation Parameters

- Temperature T: Controls randomness of sampling (0.0 nearly deterministic, above 1.0 quite random) $P(x_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$
- Top-k: Sample from k most likely tokens
- Top-p (Nucleus Sampling): Sample from the tokens whose cumulative probability $\geq p$
- Max output tokens
- Repetition / Frequency / Presence Penalty



LLM Landscape



Source: Information is beautiful

LLM Landscape

- Three big US players: OpenAI (GPT), Google (Gemini), Anthropic (Claude)
- Yes, there is also Grok (xAI)
- Chinese open-weights models
  - DeepSeek, Qwen, Kimi
- Some ~~open-source~~ open-weights models
  - LlaMA family (Meta)
  - GPT OSS (OpenAI)
  - Mistral
  - Gemma (Google)
  - Phi (Microsoft)

LLMs from Europe

- So far, the only competitive company and LLM is Mistral
  - Latest release: Mistral 3 and Mistral 3 Large (Dec 2025)
- Small endeavours from Germany and Switzerland
- There was/is one player from Heidelberg: Aleph Alpha

Braucht die deutsche Vorzeige-KI mehr Erziehung?

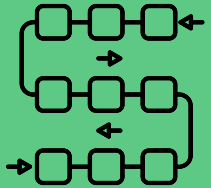
Die KI der deutschen Firma Aleph Alpha gilt als vielversprechendstes Produkt Europas. Doch sie generiert rassistische Texte. Das könnte zum Problem in Anwendungen werden.

- New server clusters are built right now (e.g. OpenEuroLLM including Jülich)
- Whether that’s fruitful and maybe even necessary?

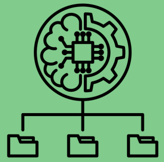
Topics that would deserve their own lecture(s)

- Efficient Adaptation
  - Parameter Efficient Fine-Tuning (PEFT)
  - Low Rank Adaptation (LoRA & QLoRA)
  - Different bit precisions
  - Adapter
- Reasoning Models
- Multimodal Capabilities
- Mixture of Experts (MoE)
- LLM Evaluations
- Explainability & Interpretability

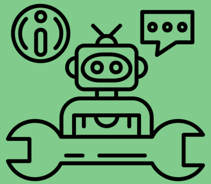
The 3 Ingredients of LLMs



Process long sequences and context



Efficient training on huge datasets



Follow (human) instructions

Tutorial