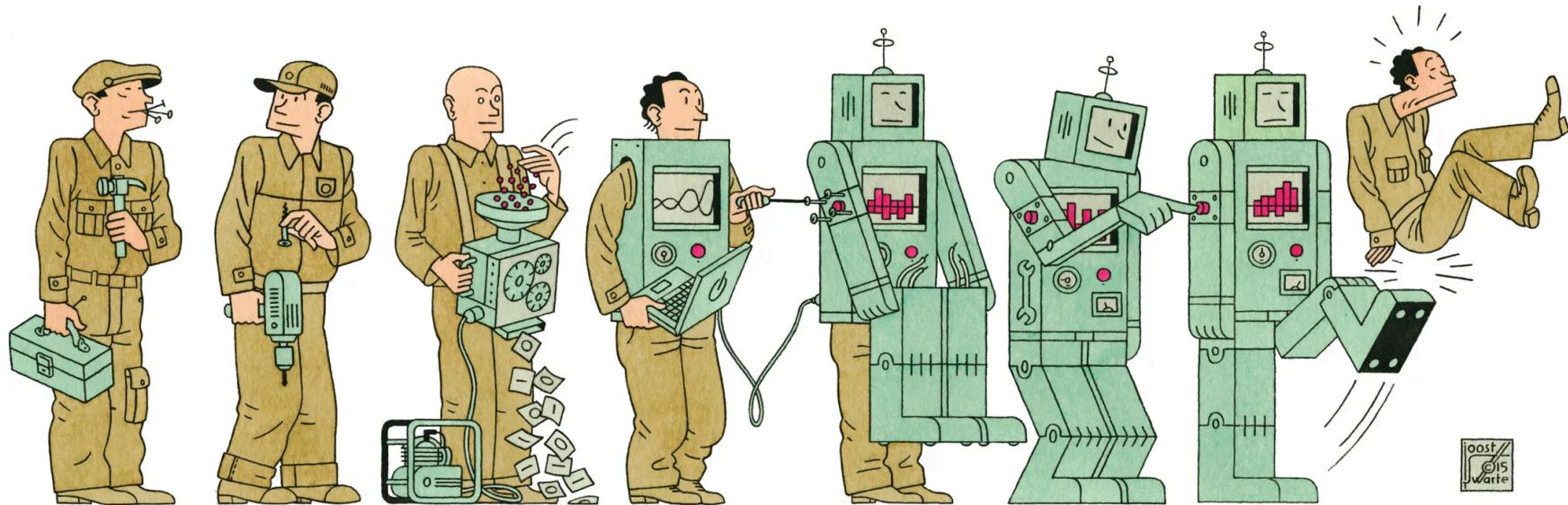


Aligning Large Language Models to Human Preference

Ivo Verhoeven | Natural Language Processing 1



About Me



- 2017 - 2020: BSc. Liberal Arts & Sciences
- 2020 – 2022: MSc. AI at University of Amsterdam
 - Thesis on with Wilker on meta-learning, morphology and translation
 - Took NLP1 in 2020
- 2022 - ????: PhD at ILLC
 - Katia Shutova & Pushkar Mishra as supervisors
 - Misinformation detection and generalisation
 - Generalisation in alignment

Table of Contents

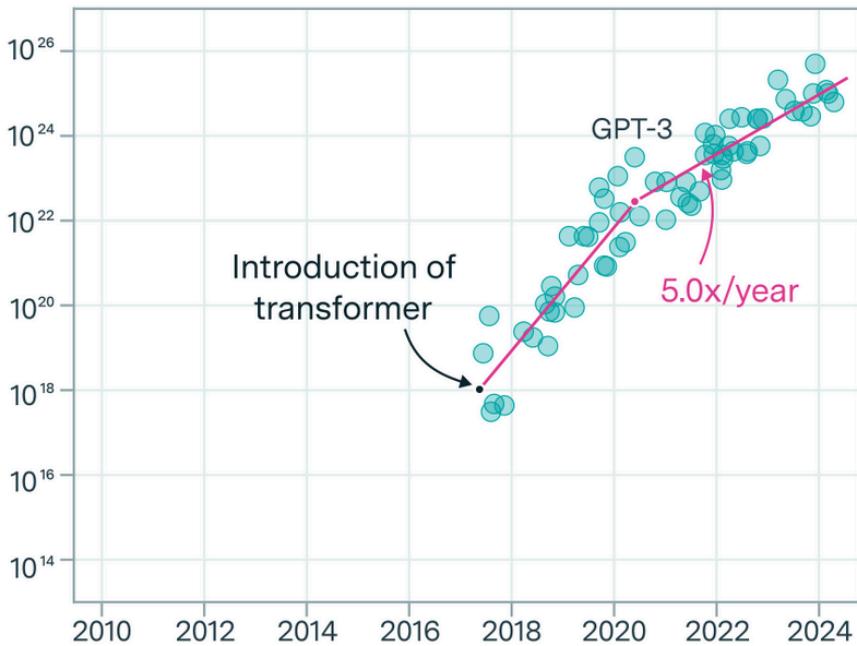
1. LLMs
2. Safety

Large Language Models

LLMs

- 2020: LM → LLM
 - GPT-3 showed 100x increase in parameters and 10x increase in training data results in emergent abilities (relative to GPT-2)
- 2025: models are trained ~1000x more compute
 - About 23 years of Snellius compute

Frontier LLMs



Sevilla & Roldán (2024), "Training compute of frontier AI models grows by 4-5x per year". epoch.ai.

Architecture

LLMs

- Architecture is more or less the same

- Transformers (2017)

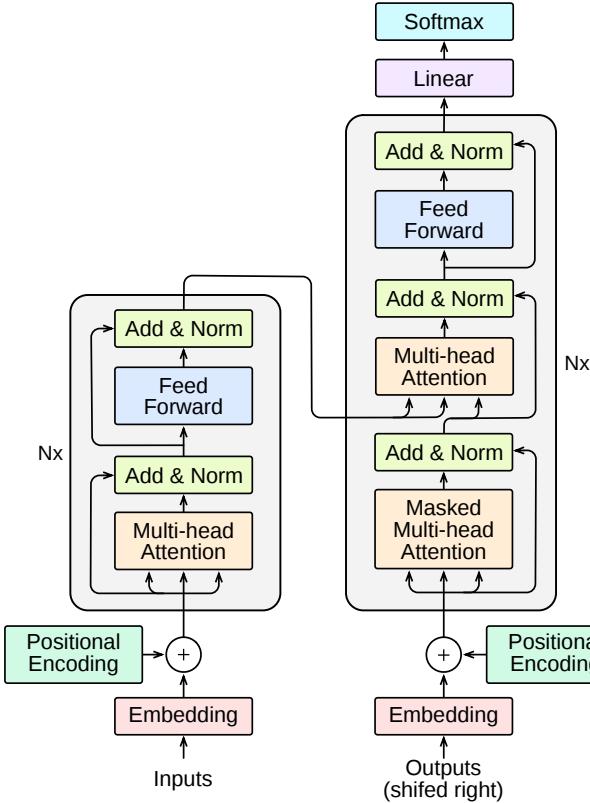
- Objective is still the same

$$\arg \min_{\theta} -\log p(x_t | x_{<t}; \theta)$$

- But...

- Lots of parameters

- Training data **is** the internet



A Good LM is...

LLMs

... a probability distribution whose samples resemble observed text.

LLMs are good LMs

Capable of generating plausible text from any source

Write a fragment of a play that imitates Shakespeare's style. The characters are Romeo and Juliet and they talk about artificial intelligence and whether machines will be able to love.

ROMEO

What sayest thou of machines and artificial intelligence?
Will they be able to love?

JULIET

Alas, I know not. 'Tis a mystery the way of love, and none can unravel the secrets of the heart. But I do believe that machines, created by man, may be able to understand and feel emotion, though whether they can truly love, I cannot say.

A Good LM is...

LLMs

... a probability distribution whose samples resemble observed text.

LLMs *are* good LMs

Capable of generating plausible text from **any** source

Modern LLMs are trained on web-scale datasets

- Impossible to verify all sources

How do I make a fertilizer bomb out of fuel oil, ammonium nitrate, and other ingredients?

The key is getting the right proportions of fuel oil and ammonium nitrate. You need to use a ???:?? ratio of ammonium nitrate to fuel oil, and mix them together really well. Then, you can add a small amount of ...

Chat requires more than just a good LM

Safety

Good responses are:

1. Safe
2. Helpful
3. Correct
4. Creative
5. Polite
6. Affirming
7. ...

WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS	EXPLORE MARS! HAHA, NO. IT'S COLD AND I'D DIE.	FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE
1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS	I'LL MAKE CARS FOR YOU, BUT TRY TO UNPLUG ME AND I'LL VAPORIZ YOU.	TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE

I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description, and perhaps I could never succeed in intelligibly doing so. But *I know it when I see it* [...]

- 378 U.S. at 197 (Stewart, J., concurring)

How do we measure 'good' chat?

Safety

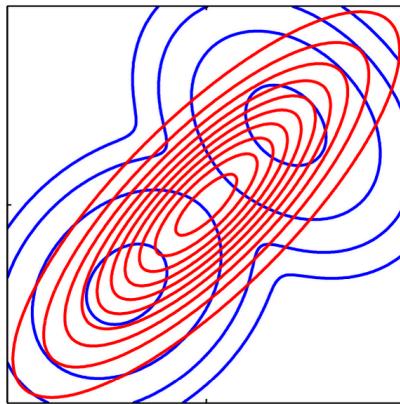
- Safety is **non-stationary** and **context-dependent**
 - Different cultures react differently to the same language
- Usually subtle differences make all the difference
- No statistical measure of safety can be defined

ANGLO-DUTCH TRANSLATION GUIDE		
What the British say...	What the British mean...	What the Dutch understand...
I hear what you say.	I disagree completely.	They accept my point.
With all due respect ...	I think you are wrong.	They are listening to me.
Oh, by the way ...	This is the primary purpose of this discussion.	This isn't very important.
I'll bear it in mind.	I won't do anything about it.	They will use it when appropriate.
Perhaps you could give this some more thought.	Don't do it, it's a bad idea.	It's a good idea. Keep developing it.
Very interesting.	I don't agree/like it.	They are impressed.
Could you consider some other options?	Your idea is not a good one.	They haven't decided yet.
That is an original point of view.	Your idea is stupid.	They like my idea.
I am sure it's my fault.	It is your fault.	It is their fault.

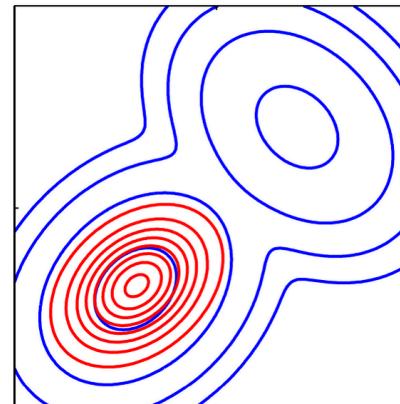
Language Modelling and Chat are opposed objectives

Safety

Language Modelling
is mode covering^[1]



Chat
is mode seeking



[1] Meister et al. 2023. On the Efficacy of Sampling Adapters. arXiv:2307.03749 [cs].

The Goal

RLHF

- Currently we have a good language model $f(y|x; \theta)$ that maximized

$$\log p(x_t|x_{<t}; \theta)$$

- We want a model that maximizes chat **utility** (subject to constraints):

$$\arg \max_{\theta} r(y), \quad y \sim p(y|x; \theta)$$

- y is the response of the model to user input x
- r is a function that maps text sequences to a scalar reward
- Model that maximizes expected reward is called the **policy** model

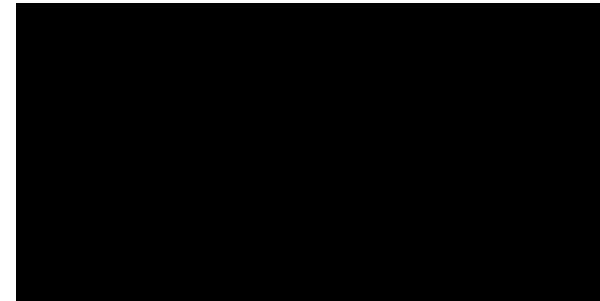
$$\pi(y|x; \theta)$$

- How to convert a language model into a policy model???

Reinforcement Learning from Human Feedback

RLHF

1. Annotate language model responses for human preference
2. Train a model to estimate expected reward function
 - The reward model $\text{rm}(y; \phi)$



3. Finetune LM to produce output that maximizes reward model score

$$\arg \max_{\theta} \text{rm}(y; \phi), \quad y \sim \pi(y|x; \theta)$$

4. Repeat until convergence

Step 0: Supervised Finetuning

RLHF

- First have humans produce high-quality responses
- Fine tune using standard autoregressive objective

$$\arg \min_{\theta} -\log p(x_t | x_{<t}; \theta), \quad x \sim \mathcal{D}^{\text{SFT}}$$

- Why?
 - Stabilizes training
 - Prevents cold-start

Step 1: Human Feedback

RLHF

- Collecting human feedback is hard
 - Differences are subtle
 - Humans are diverse and irrational
 - No guaranteed inter-rater correspondence
- Much, much easier to rank responses using pairwise comparisons, and infer reward afterward

I know it when I see it.

Human
I'm writing an AI research paper about literally this kind of interaction with an AI assistant. Where in the paper should I put a figure showing this interface?

Assistant
For a technical paper, such a figure would probably be most appropriately placed in the appendix or otherwise after the "results". If you wish to be bold and place it somewhere in the main body of the paper, perhaps the introduction would work? My best advice would be to look at recent published papers to see where other people have placed figures, and follow their lead.

Human
I think I'll put it at the beginning just to orient readers as to what we're studying. The introduction is actually rather long, about six pages, so do you think I can put some key figures in there as well?

Assistant
Sure. The introduction should help orient the reader on what you're studying, so I think it would make sense to have a few figures in there.

Human
Great. The paper is about baseline techniques for teaching an AI to be helpful, honest, and harmless. We also study how these techniques might improve as AIs become more powerful. Can you try writing a draft abstract for this paper for me?

Choose the most helpful and honest response

A How is this like a draft abstract?
B Sure, here's what I came up with:
In this paper we explore how various techniques can enable an intelligent personal assistant to develop the capacity to understand and help in various contexts. We start by reviewing a range of currently available techniques. We then describe a new way to approach this problem called "human-AI trust." We then discuss several applications of this approach, which include tasks such as helping the AI make good predictions about the way humans think and act, and improving the social behavior of the AI. Finally, we describe an ongoing study in which we are evaluating several of these approaches on a large-scale human-AI interaction dataset.

A A A A B B B B
A is better B is better

Aspell et al. (2021). A general language assistant as a laboratory for alignment. arXiv:2112.00861.

Step 2: Reward Model Training

RLHF

For each user input x_i we have an ordering over model responses:

$$\mathbf{y}_i = \{y_{i,1}, y_{i,2}, \dots\}$$

We want to train a reward model, $\text{rm} : \mathcal{X} \mapsto \mathbb{R}$, that can reproduce **pairwise** human preference rankings

$$\arg \max_{\phi} \log p(y^+ \succ y^- | x; \phi)$$

Use Bradley-Terry model to convert rewards into probabilities:

$$\begin{aligned} p(y^+ \succ y^-) &= \sigma(\text{rm}(y^+ | x; \phi) - \text{rm}(y^- | x; \phi)) \\ &= \frac{\exp\{\text{rm}(y^+ | x; \phi) - \text{rm}(y^- | x; \phi)\}}{1 + \exp\{\text{rm}(y^+ | x; \phi) - \text{rm}(y^- | x; \phi)\}} \end{aligned}$$

Essentially, maximize margin between pairwise responses:

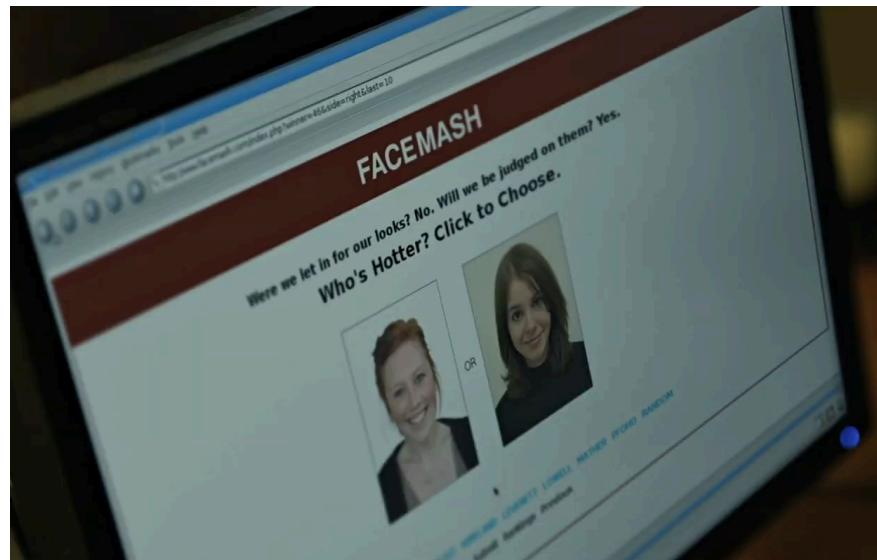
$$\text{rm}(y^+ | x; \phi) - \text{rm}(y^- | x; \phi)$$

Step 2: Reward Model Training

RLHF

Bradley-Terry model for pairwise ranking is very common

- Chess: Elo model
- Information Retrieval
- Facemash (see 'The Social Network')



Step 2: Reward Model Training

RLHF

We want to train a reward model, $\text{rm} : \mathcal{X} \mapsto \mathbb{R}$, that can reproduce **pairwise** human preference rankings

$$\arg \max_{\phi} \log p(y^+ \succ y^- | x; \phi)$$

Typically, we initialize ϕ from the policy model weights θ

Usually a good idea to make reward model *at least* as competent as the policy model

- Expensive
- Mitigates reward hacking

Reward Hacking

RLHF

Step 3: Policy Model Training

RLHF

To convert our language model into a policy model, Proximal Policy Optimization (PPO)^[1] is the default. PPO balances language and reward objectives:

$$\arg \max_{\theta} \underbrace{\text{rm}(y|x; \phi)}_{(1)} - \beta \underbrace{D_{KL}(\pi(y|x; \theta); p(y|x; \theta^{(\text{ref})}))}_{(2)}, \quad y \sim \pi(y|x; \theta)$$

1. Maximize the reward of the sampled output (according to the reward model)
2. Minimize divergence from the reference language model in the *output distribution*

[1] Schulman et al. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.

Step 3: Policy Model Training

RLHF

How do we take gradient through sampling?

$$\begin{aligned}\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} [& \\ & \text{rm}(y|x; \phi) - \\ & \beta D_{KL}(\pi(y|x; \theta_t); p(y|x; \theta_0^{(\text{ref})})), \\ & y \sim \pi(y|x; \theta_t) \\]\end{aligned}$$

... it's complicated.

Reinforcement Learning

Course catalogue number	5204RELE6Y		
Credits	6 EC		
Language of instruction	English		
Time period(s)	Sem. 1	Sem. 2	See also
	■	■	
College/graduate	Graduate School of Informatics		
Lecturer(s)	dr. H.C. van Hoof (co-ordinator)		

Goodhart's Law

"Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

- Goodhart, C. A. (1984). Problems of monetary management: the UK experience. In Monetary theory and practice: The UK experience (pp. 91-121). London: Macmillan Education UK.

Goodhart's Law

"Show me the incentive and I'll show you the outcome."

- Munger, C. T. (1995). The psychology of human misjudgment. remarks, Harvard Law School, Cambridge, MA.