Large Language Models

Implications for Linguistics, Cognitive Science & Society

Polina Tsvilodub & Michael Franke

chatGTP

envisaged by DALL-E 2 after a self-description

If you were a person, how would you look like?

As an AI language model, I do not have a physical body or appearance as I exist solely in the digital realm. However, if I were to imagine what I would look like as a person, I would imagine myself as a friendly and approachable individual, perhaps with glasses to represent my intelligence and knowledge. I would have a welcoming smile and a calm demeanor, as I strive to assist and communicate effectively with those around me. Ultimately though, since I am not a physical entity, my appearance is entirely up to the interpretation of those who interact with me.

https://www.youtube.com/watch?v=outcGtbnMuQ&t=382s

https://www.youtube.com/watch?v=outcGtbnMuQ&t=1144s

Associations, reactions, thoughts, whatever

good to withhold judgement in moral of advisory context

code pilot helpful but imperfect

weird social behavior in chat

data & privacy protection

cool tech

can be quite creative but could be entrained

human authorship

copyright

new job opportunities

old jobs endangered

problems with truthfulness

fickle in responding to different prompts

environmental impact

bad at arithmetic

challenges educational system

doesn't flag uncertainty -> bullshitting

need to know how LLMs were trained

Schedule

preliminary

session	date	topic
1	April 25	intro & overview
2	May 2	core LLMs
3	May 9	prepped LLMs
4	May 16	implications for linguistics
5	May 23	implications for CogSci
6	May 30	**holiday**
7	Jun 6	implications for society
8	Jun 13	discussion & project launch
…	…	project work
9	Jul 18	project presentations
10	Sep 1	submission deadline

Projects

Build 👷

- prompt-engineering
- LangChain agents
- generative agents
- AutoGPT applications
- RLAI fine-tuning
- ….

Test 🔎

- LLMs in the lab
  - psycholinguistics
  - CogPsy
- prompt sensitivity
- …

Create ✍🏾

- educational blog
- info video
- term paper
- survey (industry, …)
- …

Large Language Models

Core LLM

- trained on **language modeling objective**
  - predict the next word

"Here is a fragment of text … According to your **knowledge of the statistics of human language**, what words are likely to come next?

Shanahan (2022)

Prepped LLM

- trained on **usefulness objective**
  - produce text that satisfies user goals

"Here is a fragment of text … According to your **reward-based conditioning**, what words are likely to trigger positive feedback?"

Language model

left-to-right / causal model

- a **causal language model** is defined as a function that maps an initial sequence of words to a probability distribution over words: $LM : w_{1:n} \mapsto \Delta(\mathscr{V})$
  - we write $P_{LM}(w_{n+1} \mid w_{1:n})$ for the **next-word probability**
  - the **surprisal** of w_{n+1} after sequence $w_{1:n}$ is $-\log\left(P_{LM}(w_{n+1} \mid w_{1:n})\right)$
- the **sequence probability** follows from the chain rule:

$$P_{LM}(w_{1:n}) = \prod_{i=1}^{n} P_{LM}(w_i \mid w_{1:i-1})$$

- measures of **goodness of fit** for observed sequence $w_{1:n}$:
  - **perplexity**:

$$\mathrm{PP}_{LM}(w_{1:n}) = P_{LM}(w_{1:n})^{-\frac{1}{n}}$$

  - **average surprisal**:

$$\text{Avg-Surprisal}_{LM}(w_{1:n}) = -\frac{1}{n}\log P_{LM}(w_{1:n})$$

$$\log \mathrm{PP}_{M}(w_{1:n}) = \text{Avg-Surprisal}_{M}(w_{1:n})$$

Self-attention layer

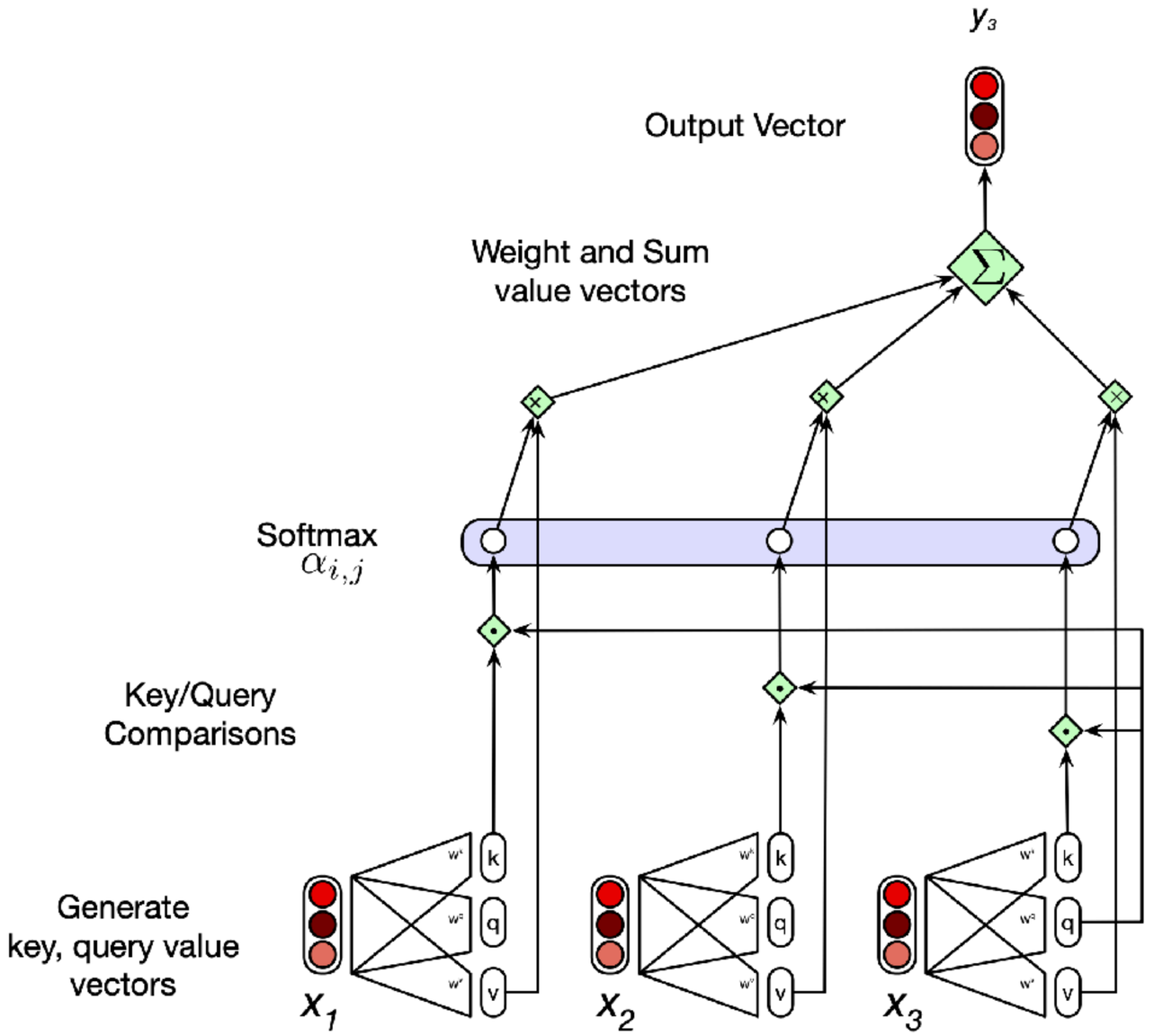
- **output**

$$\mathbf{y}_i = \sum_{j \le i} \alpha_{ij} \mathbf{v}_j$$

- **weight score**

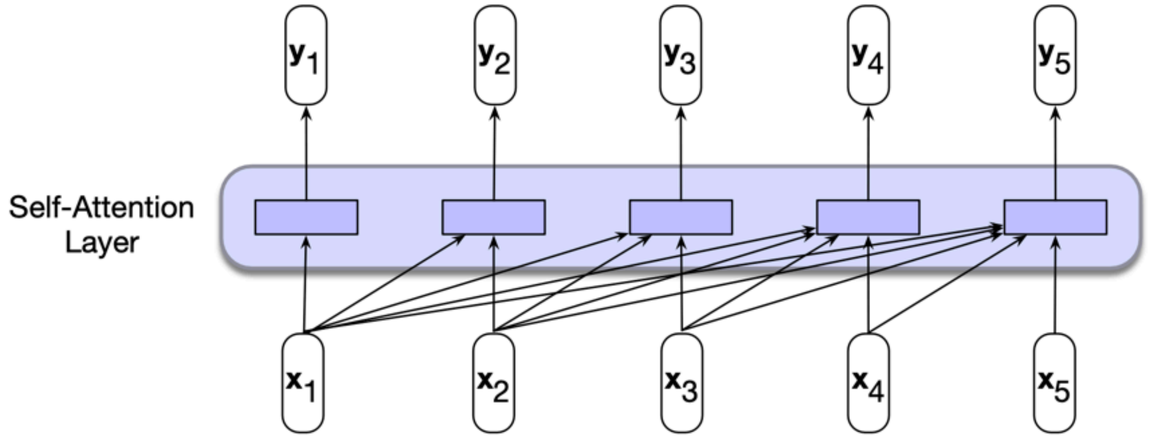
$$\alpha_{i,j} = \frac{\exp(\mathbf{q}_i \cdot \mathbf{k}_j)}{\sum_{j' \le i} \exp(\mathbf{q}_i \cdot \mathbf{k}_{j'})}$$

- three vectors for each input vector x_i
  1. **query:** which info to extract from context
     $\mathbf{q}_i = \mathbf{W}^Q \mathbf{x}_i$
  2. **key:** which info to provide for later
     $\mathbf{k}_i = \mathbf{W}^K \mathbf{x}_i$
  3. **value:** what output to choose
     $\mathbf{v}_i = \mathbf{W}^V \mathbf{x}_i$



Vaswani et al. (2017)

Causal LM

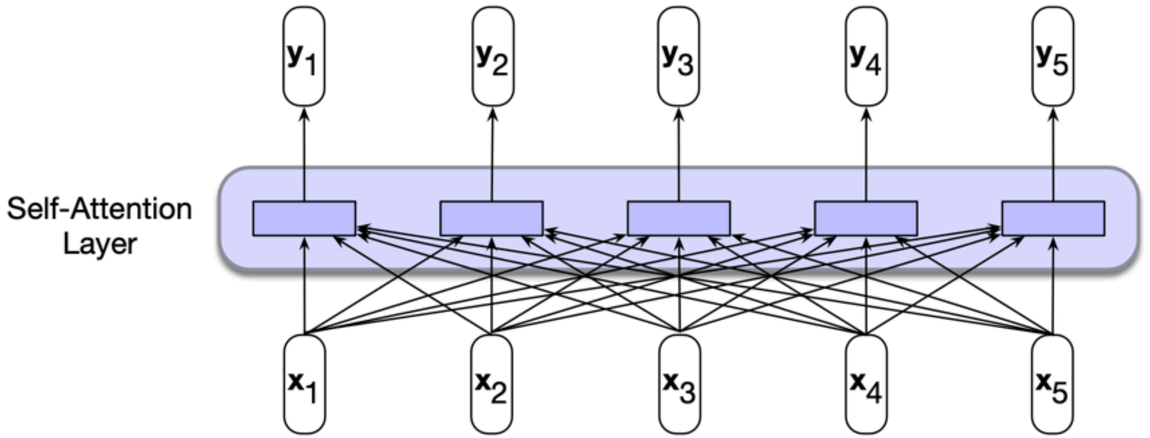


computation for input $\mathbf{x}_1, \ldots, \mathbf{x}_3$ blind to $\mathbf{x}_4$ and $\mathbf{x}_5$

$\mathbf{y}_5$ is embedding for input $\mathbf{x}_1, \ldots, \mathbf{x}_5$

$\mathbf{y}_5$ is a “left-contextual embedding”

Bidirectional encoder



computation for input $\mathbf{x}_1, \ldots, \mathbf{x}_3$ sees $\mathbf{x}_4$ and $\mathbf{x}_5$

$\mathbf{y}_1, \ldots, \mathbf{y}_5$ is embedding for input $\mathbf{x}_1, \ldots, \mathbf{x}_5$

$\mathbf{y}_i$ are bidirectional “contextual embeddings”

Prepped LLMs

Fine-tuning and RLHF / RLAI

- in certain contexts, we might not want to generate the *most likely* next words
  - follow instructions
  - useless or impolite responses, toxic language
  - code for illegal activities
  - …
- to fix this, fine-tune the model to satisfy the users' preferences via **reinforcement learning from human feedback**
  - incentivise the *agent* with a *reward* when its output matches achieves the *goal:*

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

  - adjust the *policy* so as to maximize the expected *return*:

$\pi(s_t) = P(a_t \mid s_t)$ and adjust policy to maximize $L_\theta = \mathbb{E}_t[G_t \; log \; \pi_\theta(a_t \mid s_t)]$

  - formulate the *reward* function based on comparative preferences

Sutton & Barto (2018), Bai et al (2022)

RLHF

Reinforcement Learning from Human Feedback

- used to fine-tune e.g. GPT-3.5, GPT-4 and ChatGPT (Brown et al., 2020; OpenAI, 2022)

Image: https://openaicom.imgix.net/cf717bdb-0c8c-428a-b82b-3c3add87a600/ChatGPT_Diagram.svg?fm=auto&auto=compress,format&fit=min&w=1919&h=1138

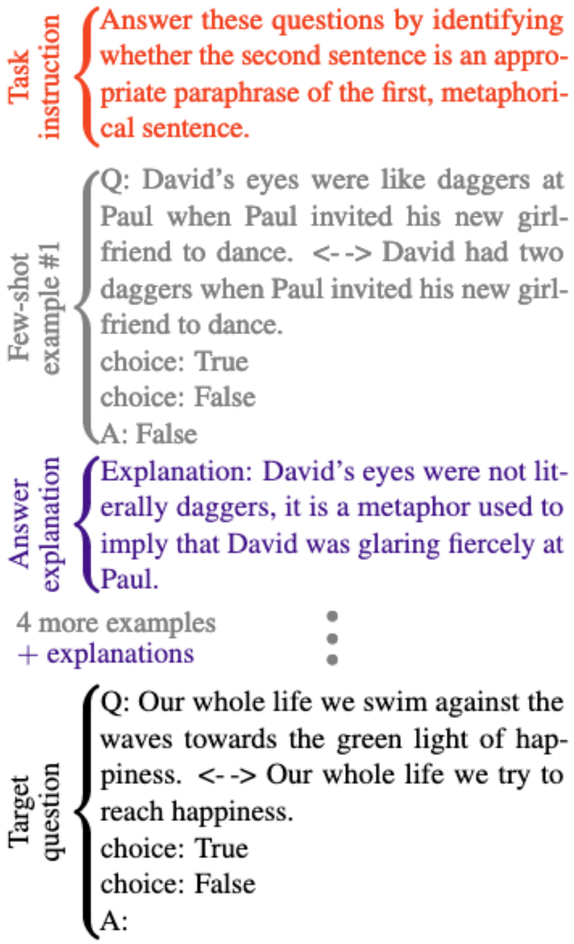
- RM: fine-tuned GPT-3 (6B in InstructGPT) trained to output scalar reward for prompt x and completion y_w (preferred over y_t)
- RM is used to train the LLM via RL
- policy trained via *proximal policy optimization* (PPO) with bells and whistles

Christiano et al. (2017), Ouyang et al. (2022), Stiennon et al. (2022)

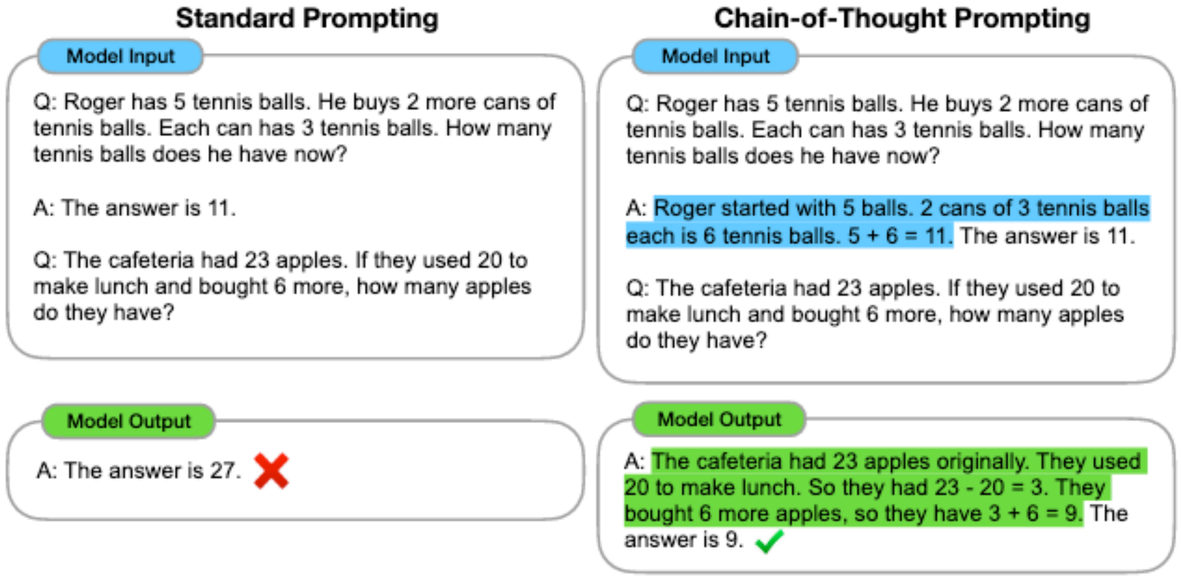
Prompting LLMs

Few-shot and Chain-of-Thought

- the users might want to adjust the model output to their particular needs



- the model might need "working memory" to solve the task



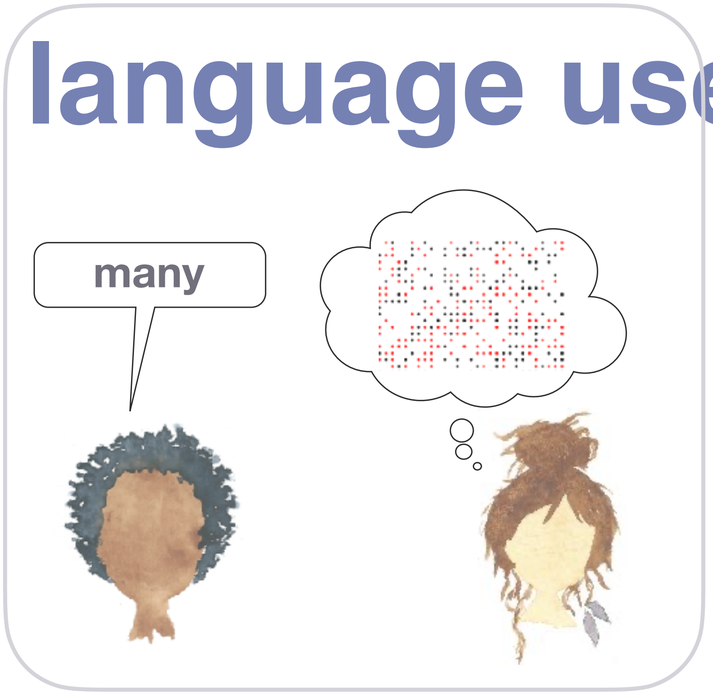
Wei et al (2022), Lampinen et al (2022)

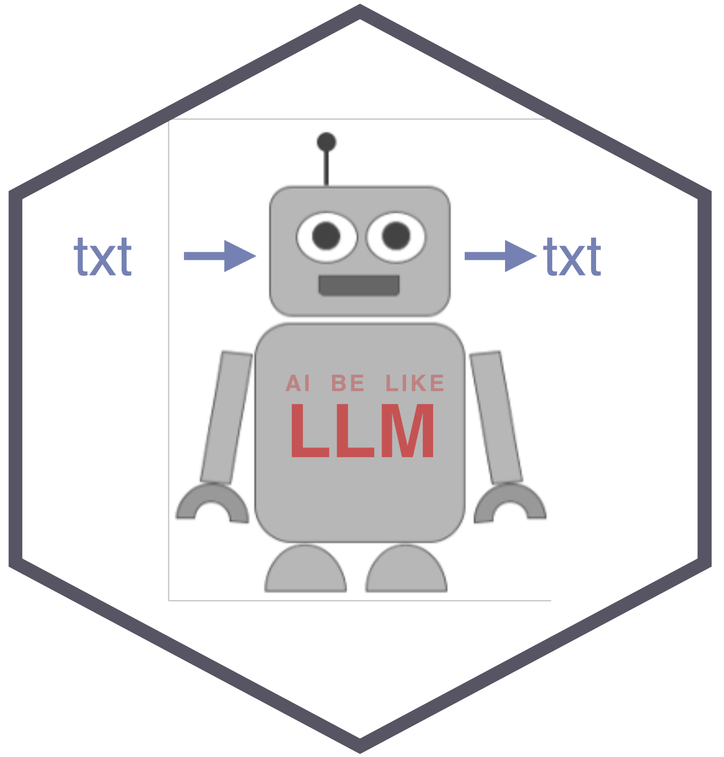
Prompting LLMs
Instructions

- LLMs are (among other things) fine-tuned to follow instructions
- instruction following opens up an avenue for a vast space of functions the model will perform
  - Q: …. A: ….
  - Write Java code for X
  - Edit X to be Y
  - Here is tool X and how it works, reason step by step and decide when to use it for solving task Y
  - Here is a list of tools, decide which of them to use for task X
  - …

Linguistics

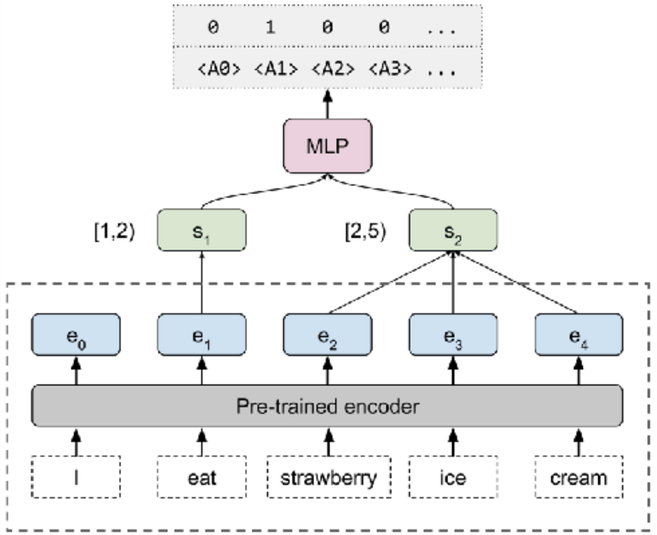
Language: solved!



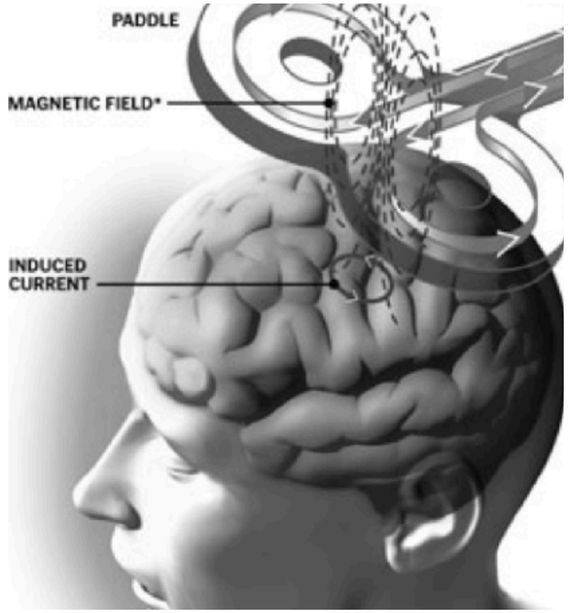


Chartering “linguistic knowledge” of LLMs

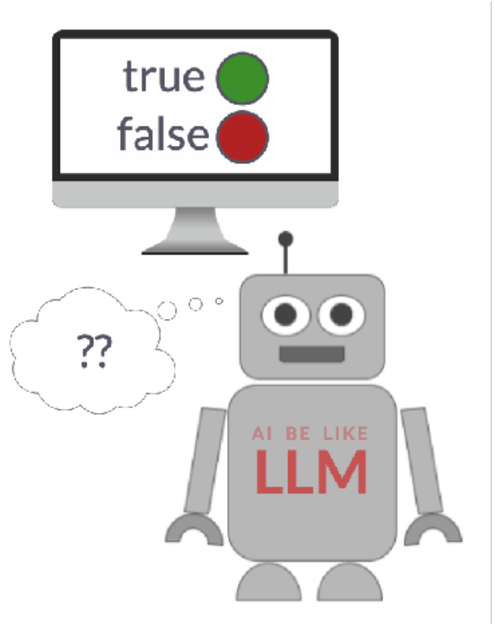
Probing



Intervention



Behavior. Tests



NLP Benchmarks

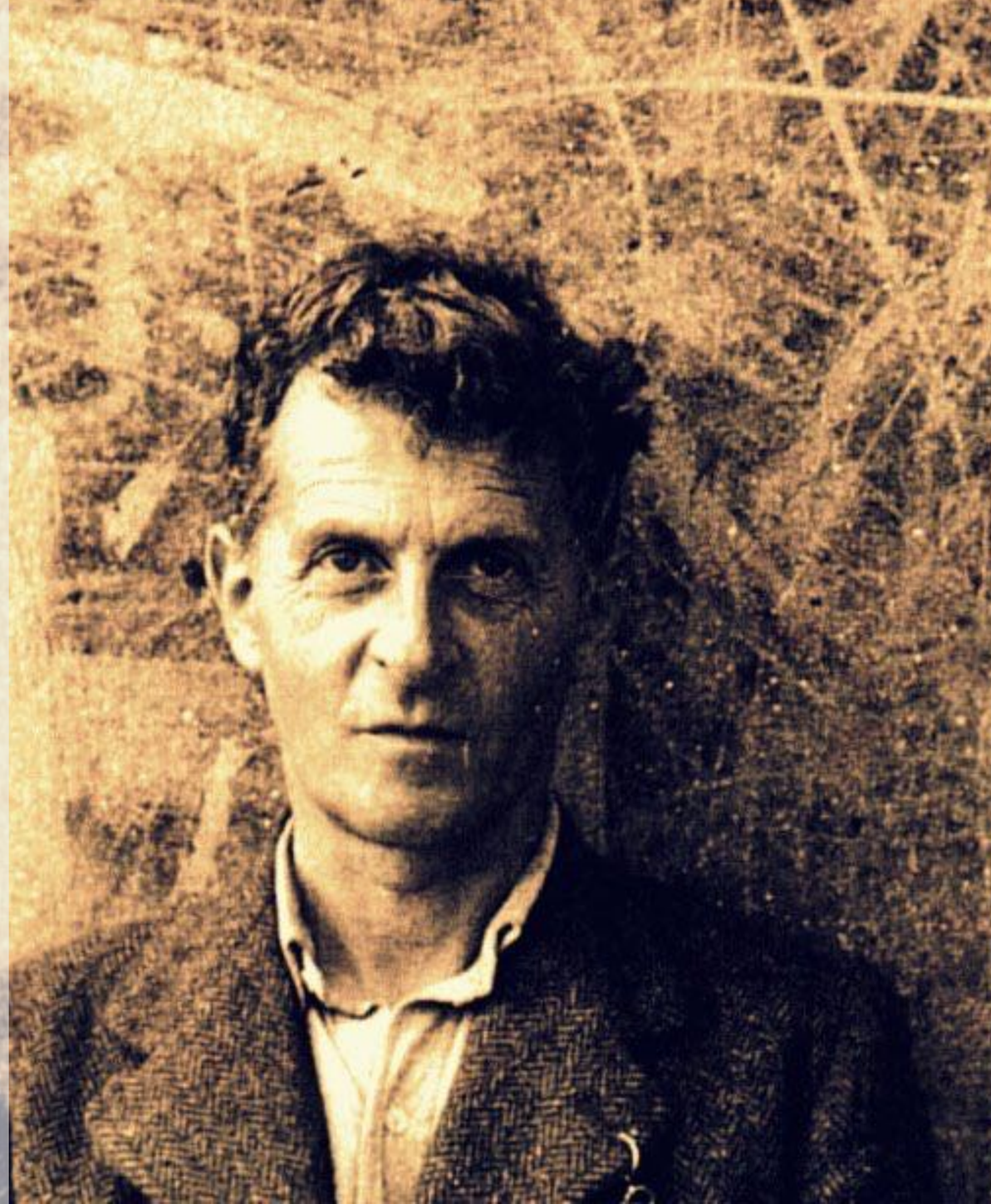
Quantifying LLM intelligence

- testing linguistic knowledge
  - MNLI, SuperGLUE (semantics), COLA, LAMBADA (long-distance dependencies), ImpPres (pragmatics),…
- testing reasoning abilities
  - math: GSM8K, SVAMP,…
  - common sense: StrategyQA, HellaSwag,…
- testing factual knowledge
  - question answering: TriviaQA.…
  - reading comprehension: RACE,…
- misc: bar exam, SATs, HumanEval (coding),…
- testing biases: WinoGrande, BBQ
- [benchmarks 2.0] generated by LLMs for LLMs (Perez et al, 2022)
  - evaluating personas (‘world views’, agreeability,…), sycophancy, safety

BLEU
METEOR
ROUGE

Cognitive Science
& Philosophy of Mind

Einen Satz verstehen heißt, wissen, was der Fall ist, wenn er wahr ist.



Understanding understanding

1. Do LLMs understand language?
   Depends on what it means to understand language.

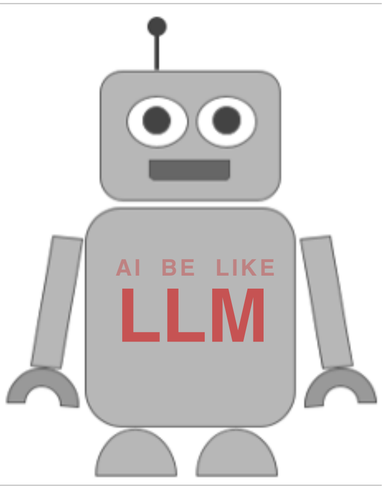
2. Do LLMs understand the world?
   Depends on what it means to understand the world.

3. How can we understand how LLMs work?
   Depends on whether the LLM wants us to understand.

Two forms of intelligence

or: the LLM cheat sheet

NEITHER OF WHICH ANYONE REALLY FULLY UNDERSTANDS



language
(written)

user feedback

statistical regularities between words (tokens)

language
(written)

language
(written, spoken, signed)

world
(the whole gory mess)

abstract, generic, & causal concepts

language
(written, spoken, signed)

action
(all the fails & glory)

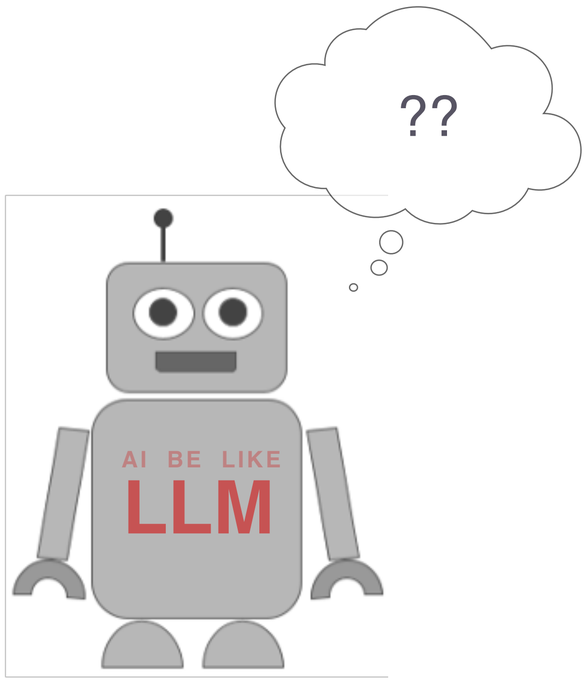
All penguins are black & white.

Some old TV shows are black & white.

Therefore, all penguins are old TV shows.

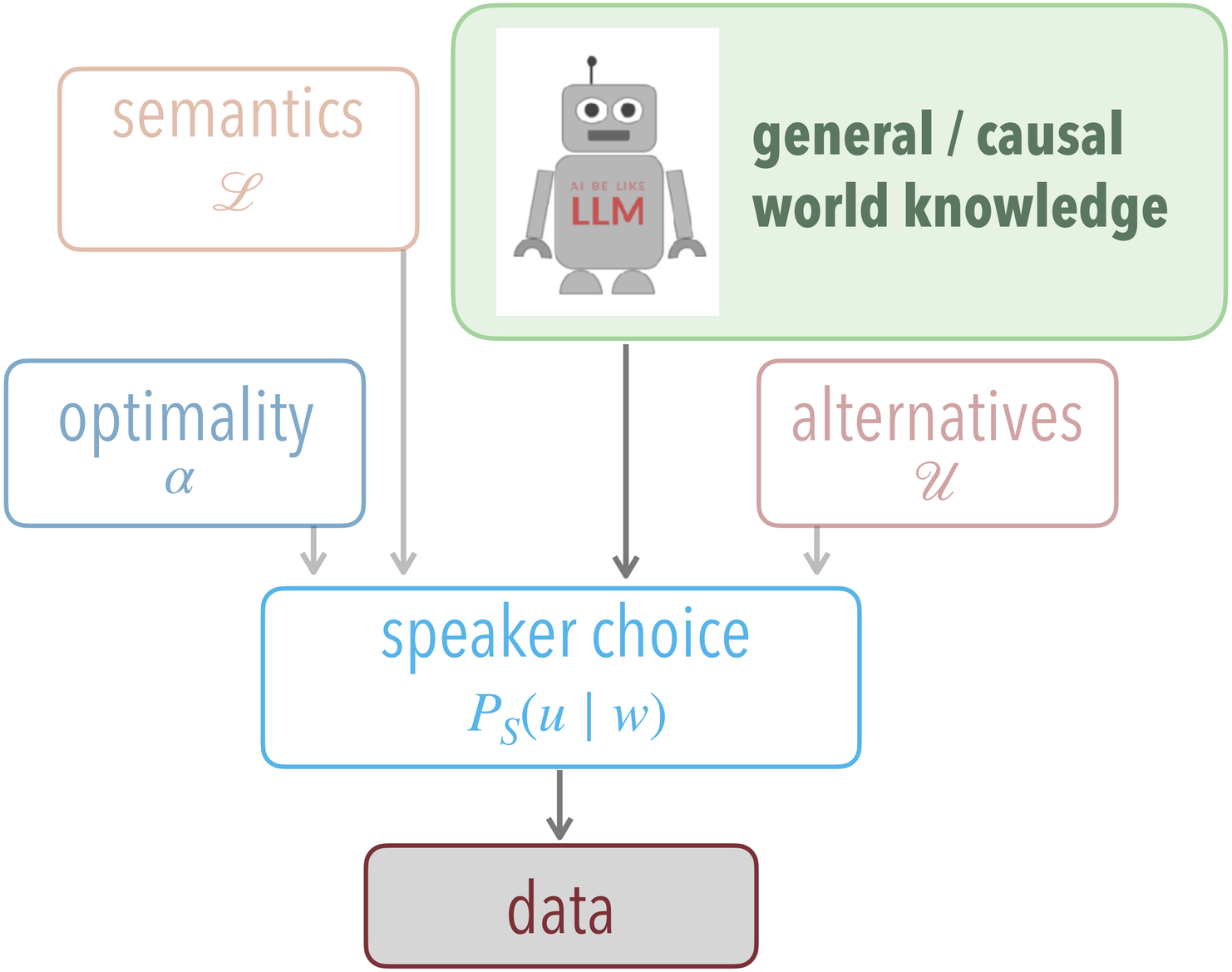
valid

invalid



Hybrid cognitive models

integrating LLMs in explanatory models



Society & Ethics


De grootste verzameling paperclips is in het Noord-Hollandse Spaarnwoude.

Ethical considerations

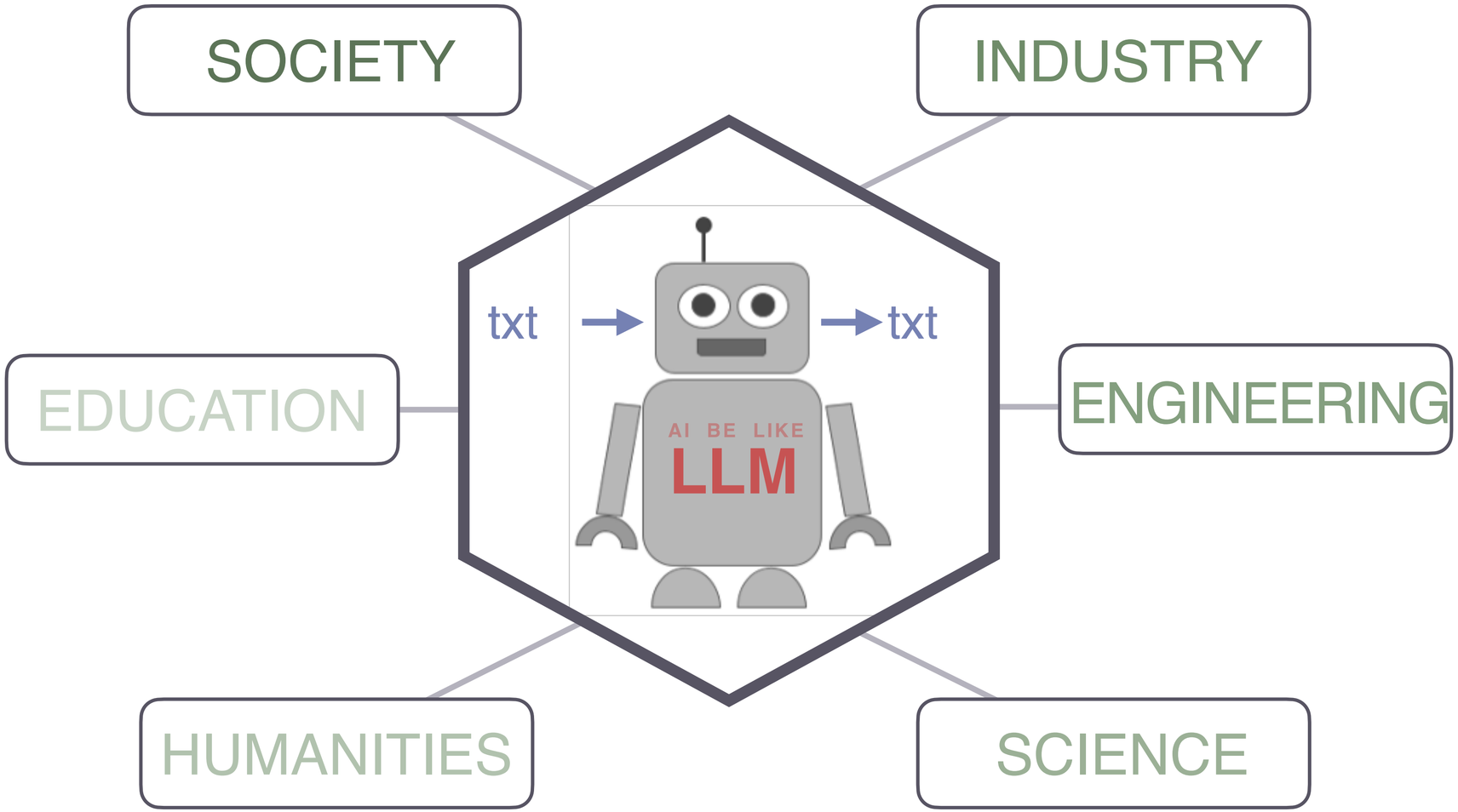
example: stereotypes

from here

WEIRD WYOMING

- just as experimental psychology is **WEIRD**
  - Western
  - Eductated
  - Industrialized
  - Rich
  - Democratic
- usual LLM training data is from **WYOMING**
  - Western
  - Young
  - Opinionated
  - Males with
  - Internet from
  - Non-marginalized
  - Groups

Bender et al. (2021)


SOCIETY
INDUSTRY
txt
txt
AI BE LIKE
LLM
EDUCATION
ENGINEERING
HUMANITIES
SCIENCE