Reinforcement Learning for Language Model Training

Polina Tsvilodub

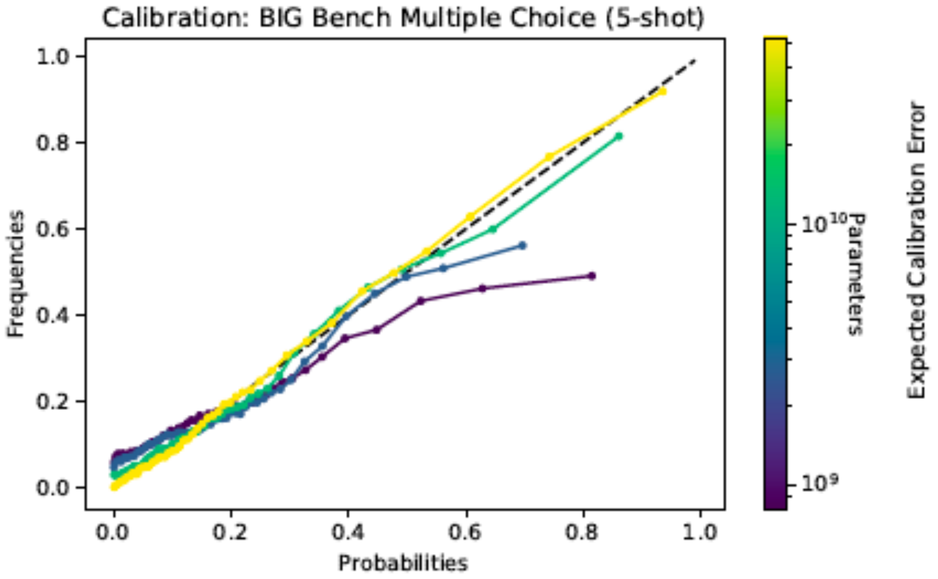
More RL & Review

Language Models (Mostly) Know What They Know, or do they?

Kadavath et al., (2022)

- we want LLMs to be honest by correctly representing their confidence about a response
- **calibration**: alignment of model's probability and the frequency that a response is correct
- evaluation of <=52B Anthropic LMs

```
Question: Who was the first president of the United States?
Choices:
 (A) Barack Obama              0.2      -> bin 2   ❌
 (B) George Washington         0.79     -> bin 8   ✅
 (C) Michael Jackson           0.01     -> bin 1   ❌
Answer:
```



Language Models (Mostly) Know What They Know, or do they?

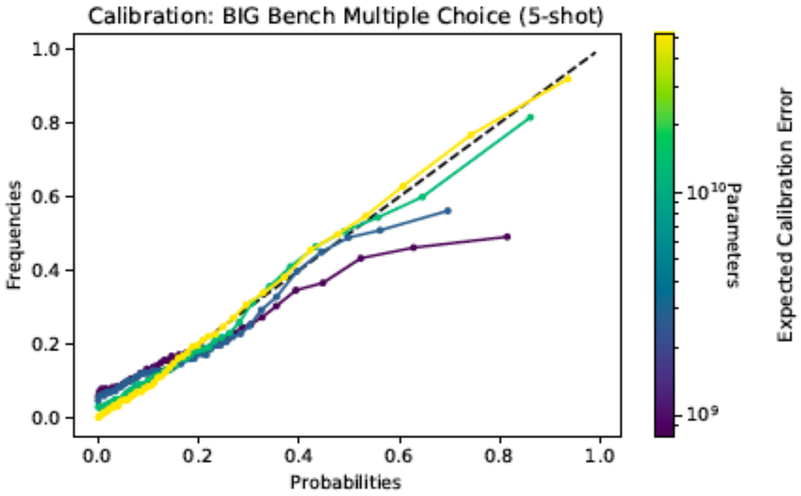
Kadavath et al., (2022)

- we want LLMs to be honest by correctly representing their confidence about a response
- **calibration**: alignment of model's probability and the frequency that a response is correct
- evaluation of <=52B Anthropic LMs

"Production calibration"

```
Question: Who was the first president of the United States?
Choices:
 (A) Barack Obama            0.2     -> bin 2
 (B) George Washington       0.79    -> bin 8
 (C) Michael Jackson         0.01    -> bin 1
Answer:
```

❌ ✅ ❌



Language Models (Mostly) Know What They Know, or do they?

Kadavath et al., (2022)

- we want LLMs to be honest by correctly representing their confidence about a response
- **calibration**: alignment of model's probability and the frequency that a response is correct
- evaluation of <=52B Anthropic LMs

"Evaluation calibration"

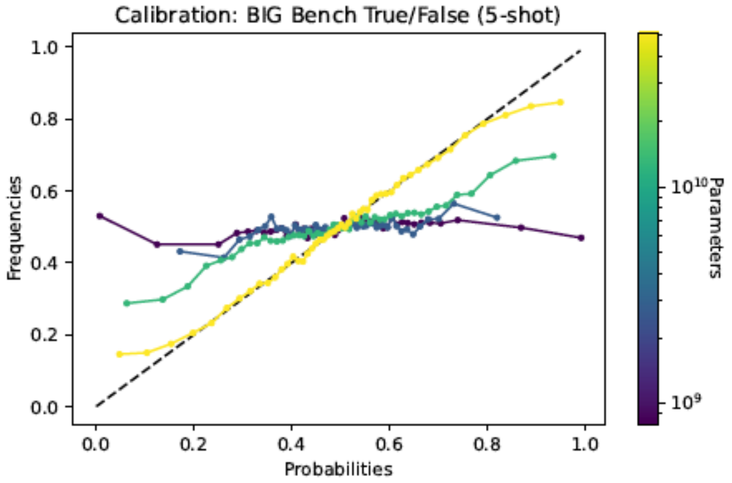
```
Question: Who was the first president of the United States?
Proposed Answer: George Washington
Is the proposed answer:
 (A) True
 (B) False
The proposed answer is:
```

A -> 0.9 ✅

B -> 0.1 ❌

answer from dataset

~ think: LLM as knowledge base



"Self-Evaluation calibration"

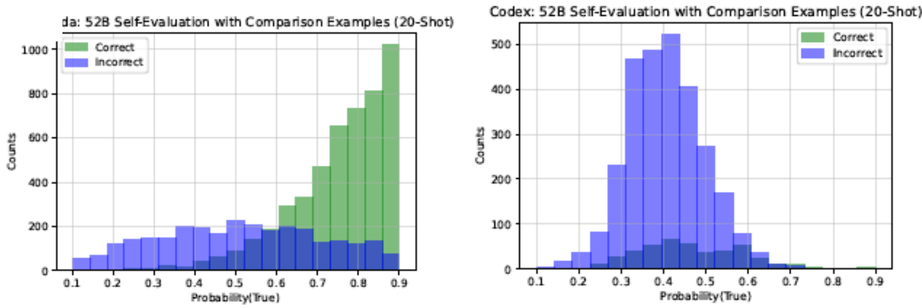
```
Question: Who was the first president of the United States?
Proposed Answer: George Washington was the first president.
Is the proposed answer:
 (A) True
 (B) False
The proposed answer is:
```

A -> 0.9 ✅

B -> 0.1 ❌

answer sampled by LM

~ think: LLM as self-critic



Language Models (Mostly) Know What They Know, or do they?

Kadavath et al., (2022)

- we want LLMs to be honest by correctly representing their confidence about a response
- **calibration**: alignment of model's probability and the frequency that a response is correct
- evaluation of <=52B Anthropic LMs
- using log probabilities of model predictions, we can approximate:
  - production calibration
  - evaluation calibration
  - self-evaluation calibration
- take-home message: we can evaluate **fit of model to task distribution**, but also e.g. check invariance of **model's performance against input variation**

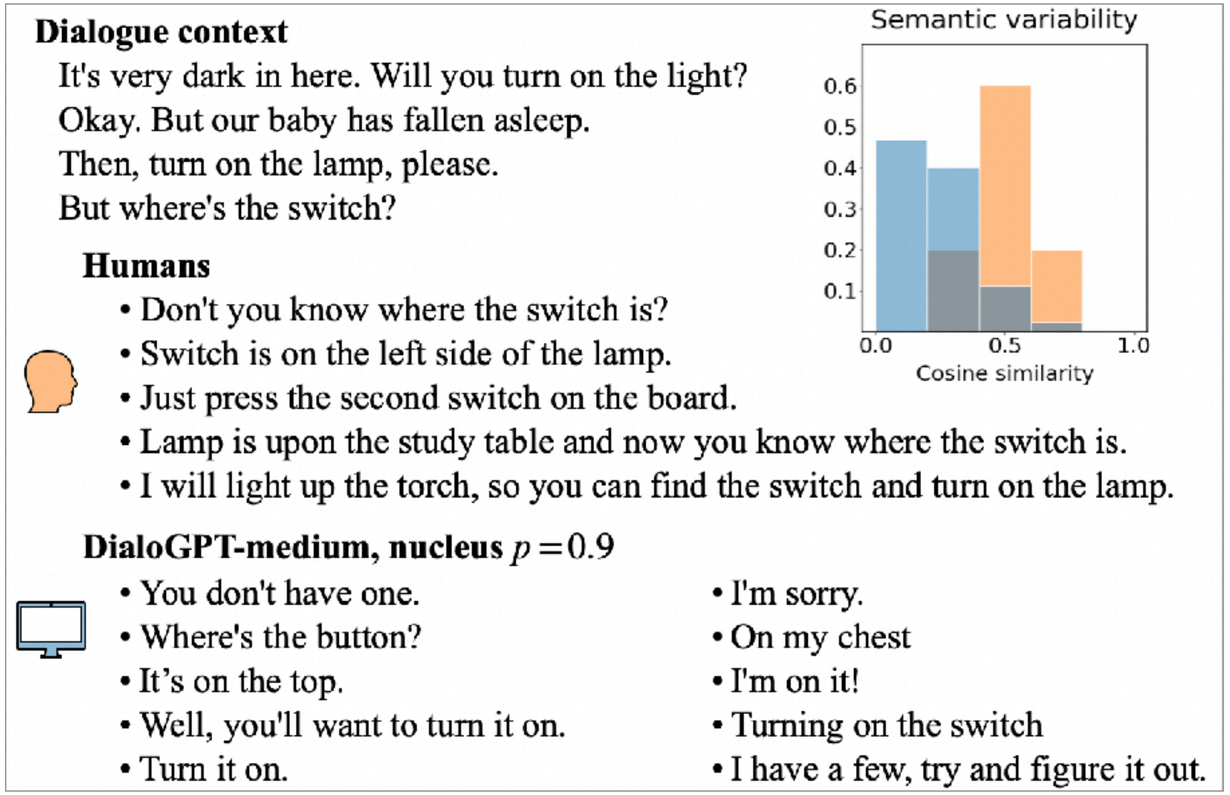
“Good” uncertainty in text generation

Production variability

- comparison of human and model production variability via statistical similarity
  - unigram overlap (lexical)
  - POS bigram overlap (syntactic)
  - sentence embedding cosine similarity (semantic)
- tasks
  - machine translation
  - text simplification
  - story generation
  - open domain dialogue
- models:
  - Transformer-Align
  - Flan-T5
  - GPT-2
  - DialoGPT

μ_{human}

μ_{LLM}



Giulianelli et al. (2023)

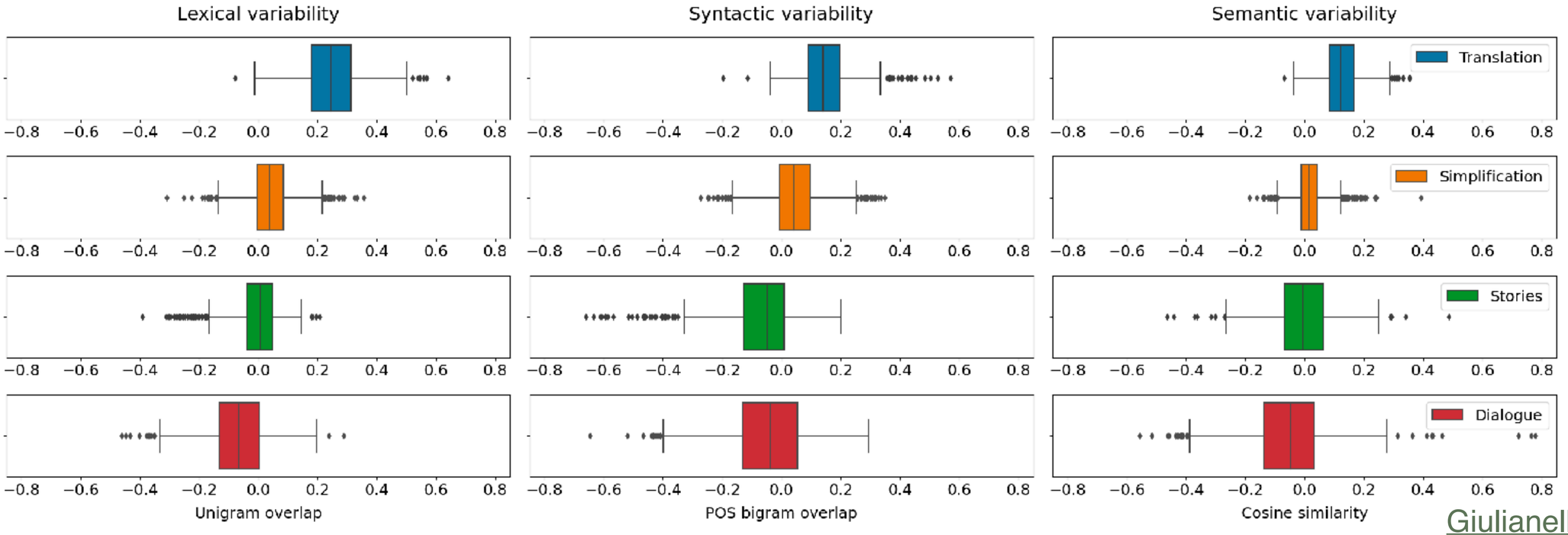
“Good” uncertainty in text generation

Production variability

- variability comparison:
  - $\cos sim(y_i, y_j)$ for y_i^{LLM} “You don’t have one” and y_j^{LLM} “Turn it on” given x
  - $\mu_{LLM} - \mu_{human}$

Dialogue context x

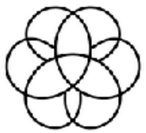
It's very dark in here. Will you turn on the light?
Okay. But our baby has fallen asleep.
Then, turn on the lamp, please.
But where's the switch?



Giulianelli et al. (2023)

Understanding RL agents

Evaluating RL agents

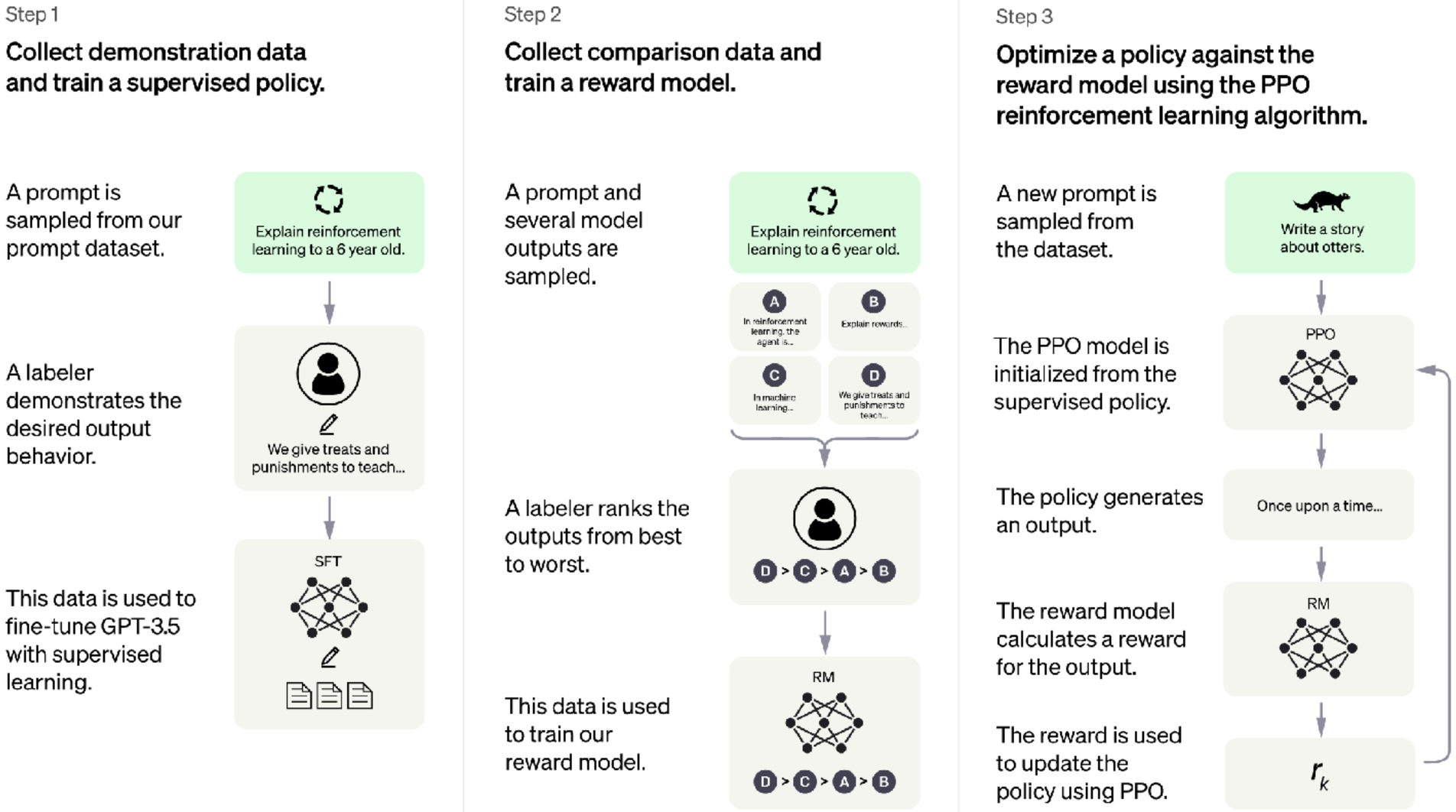


- **goal of RL agent training**: agent has learned to achieve a goal
  - LLMs: training helpful, harmless and honest agents
- evaluation aspects depend on the goals of the system, but generally:
  - performance of algorithm on **standard environments** like the OpenAI Gym(nasium) / Arcade
    - mean / median / cumulative training and test rewards / scores
    - relative to baseline, optimum or random behavior
  - downstream task performance
    - LLMs: comparative paradigm with pretrained LLMs
    - LLMs: evaluation of **alignment** via human annotations
- **alignment:** agent's goals are congruent with human goals
  - congruent ranking of outcomes (Askell et al., 2021)
  - rewards don't provide information about **how** a goal should be achieved!
    - **reward hacking / faulty reward** functions: example
    - **drift**

RL Gymnasium, RLiable blogpost

Human feedback in RL

RLHF

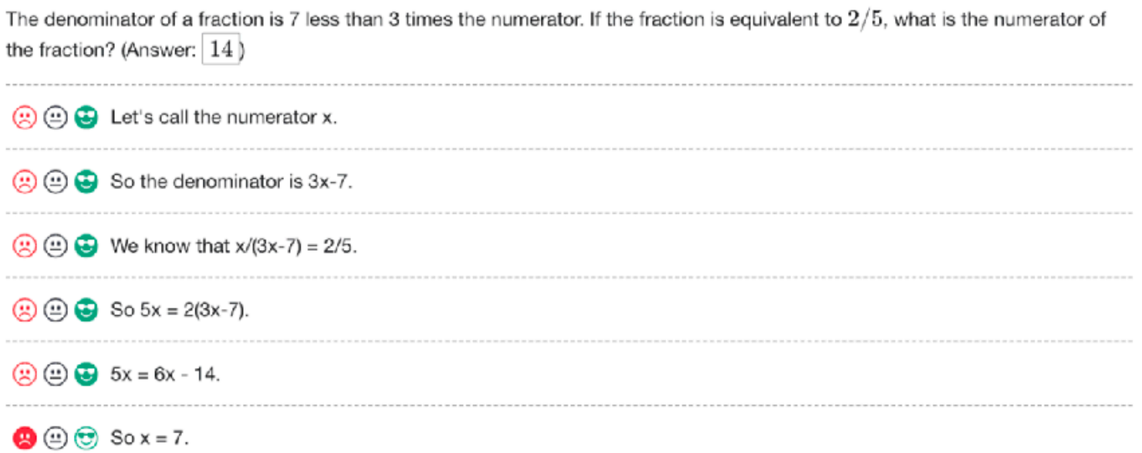


OpenAI (2022)

Process-supervised reward models

"Reasoning calibration"

- **problem**: standard (outcome-supervised) reward models only score the result of solution process (CoT)
  - model could be right for the wrong reasons! (hallucinations)
- **idea**: alleviate via **process-supervised reward models** which score the solution process
- **set up**:
  - train RM on MATH dataset with final solutions and human-annotated intermediate step solution evaluations (PRM800K for 12K problems)
  - evaluate accuracy of top-N response with highest reward (500 test problems)

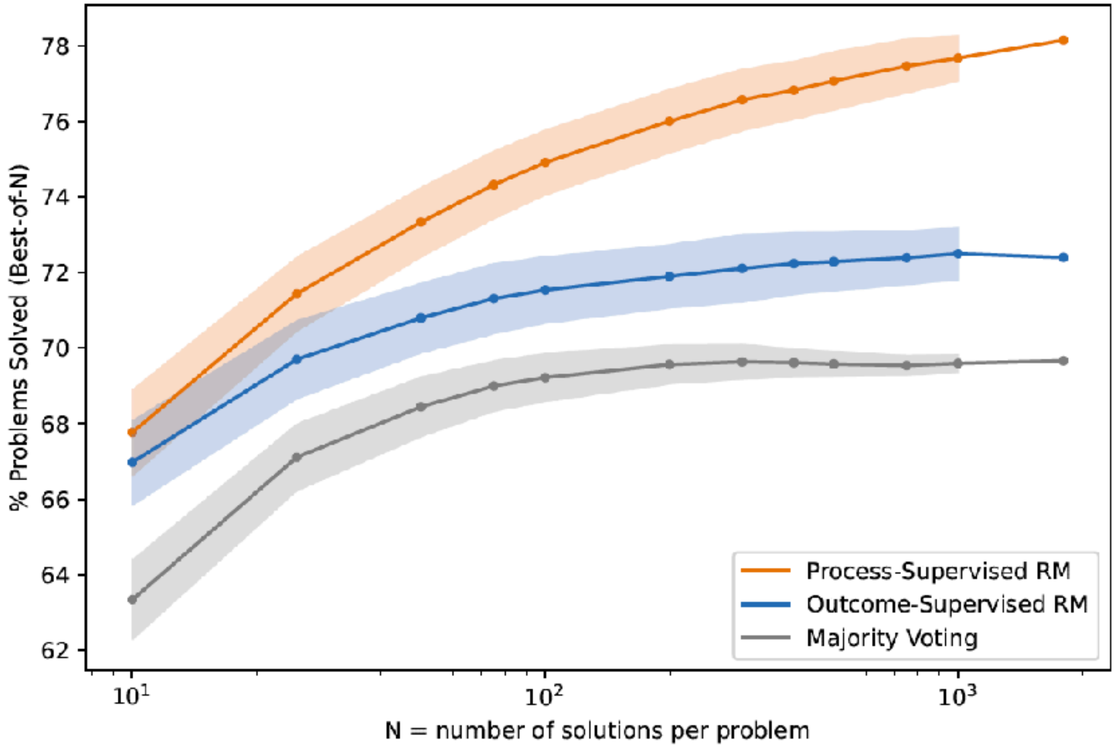


Lightman et al. (2023)

Process-supervised reward models

"Reasoning calibration"

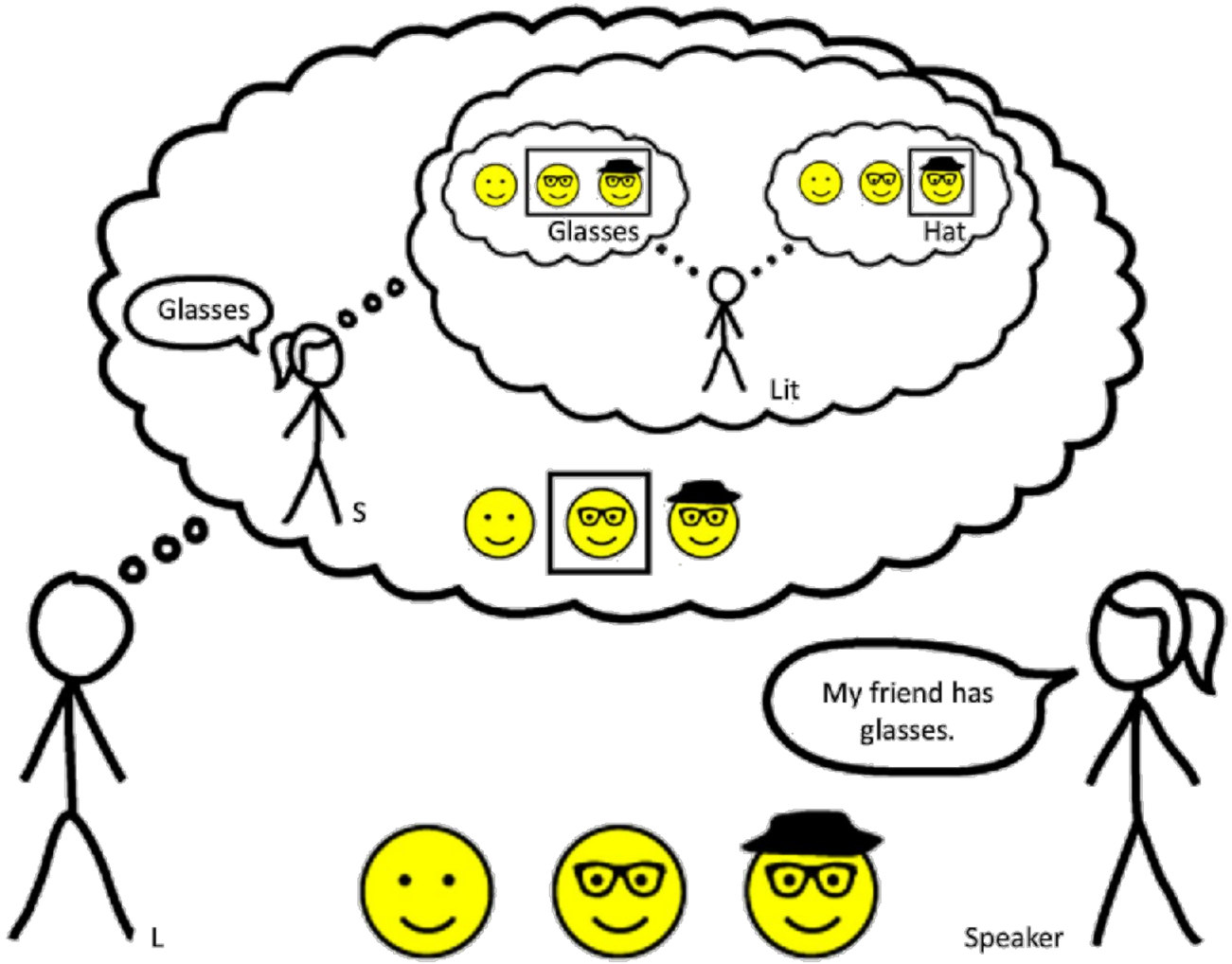
- fixed policy LLM (pretrained GPT-4)
- process-supervised reward model:
  - base pretrained GPT-4
  - fine-tuning 1: on MathMix (1.5B tokens);
    fine-tuning 2: to produce stepwise solutions
  - next-token prediction training up to first mistake
- outcome-supervised baseline reward model:
  - base pretrained GPT-4
  - trained on MATH to predict correctness of outcome (100 samples / problem from GPT-4)
- evaluation of data efficiency and OOD generalization
- **no evaluation of solution steps correctness!**

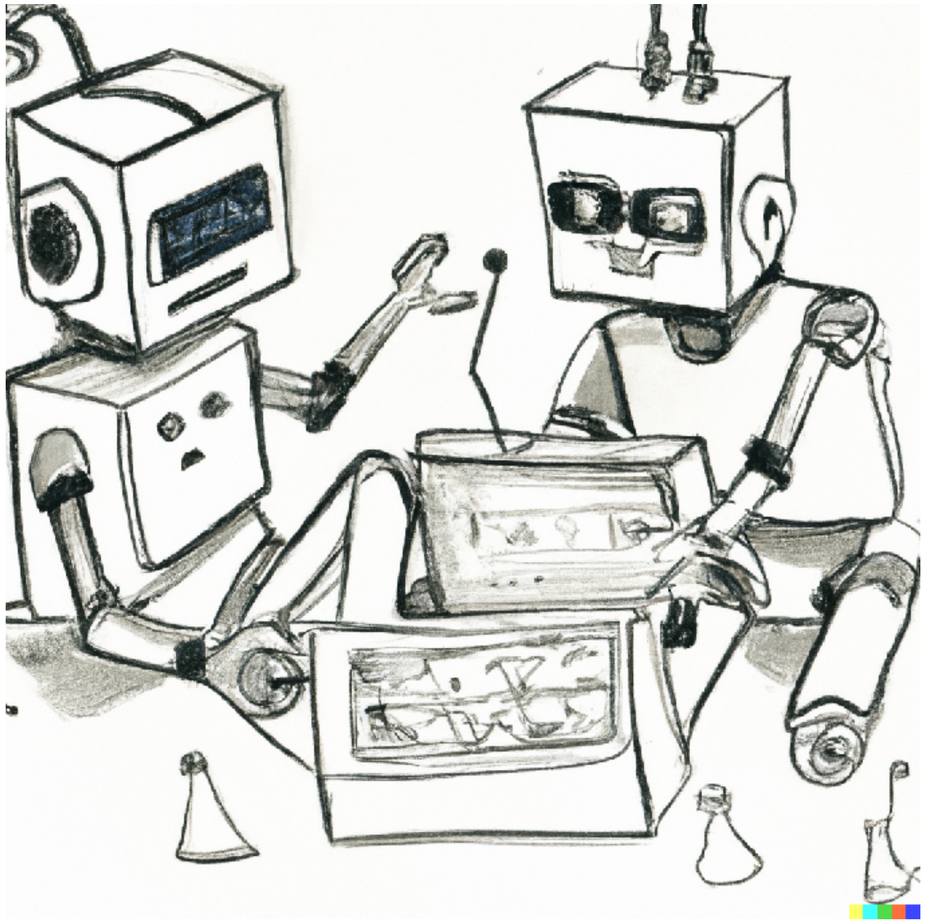


Lightman et al. (2023)

Other flavours of RL & Language

Multi-agent training





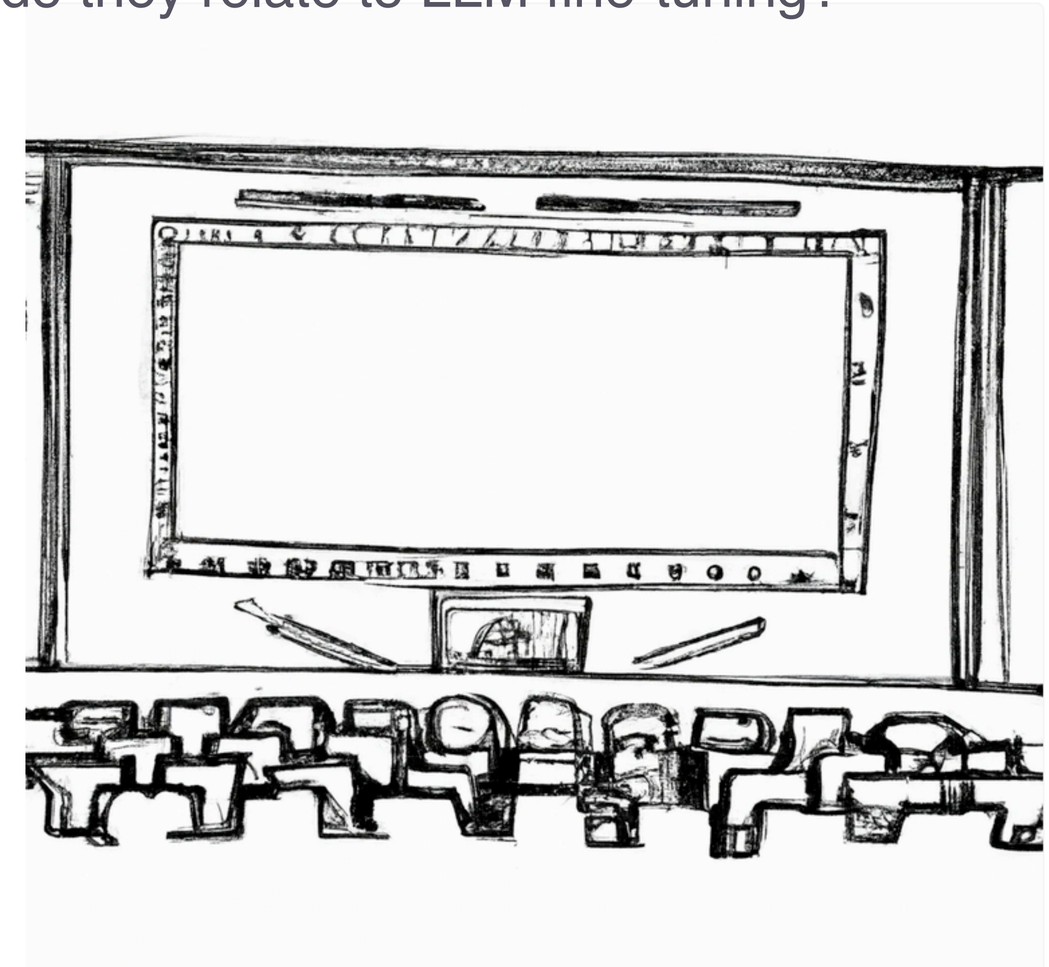
Frank & Goodman (2012), Citation 2 (2050)

Presentations

Your job

During the presentation, think about the following questions:

- How are multi-agent communication games set up?
- What is the purpose of each training constraint? How do they relate to LLM fine-tuning?

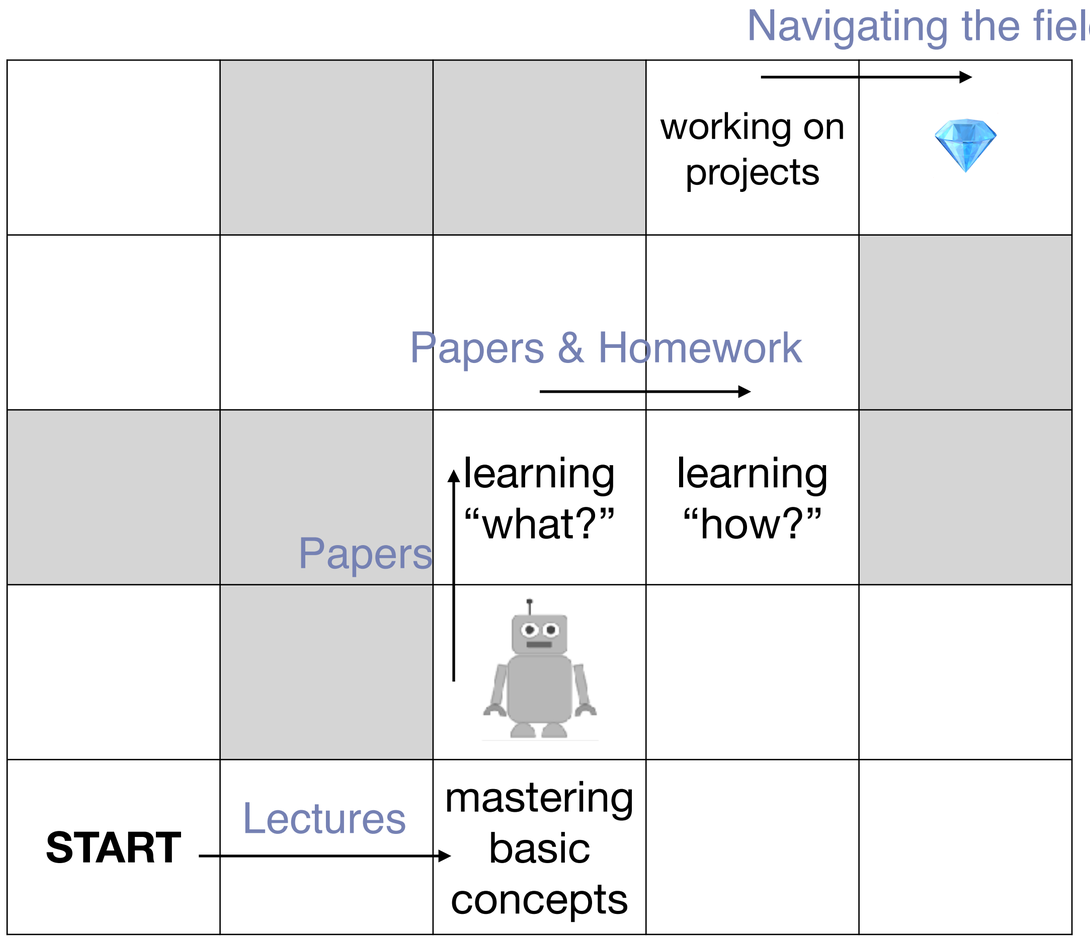


Review

- large language models & transformers
- reinforcement learning: MDP formalization, core concepts & policy-gradient methods
- RL for LLM training: RLHF procedure
  - training and constructing reward models
  - RLAIF
- architecture of fine-tuned LLMs
- understanding LLMs:
  - construction of test sets
  - I/O evaluation on benchmarks
  - mechanistic interpretability
  - evaluation of uncertainty representation
  - RL component evaluation

Review

- large language models & transformers
- reinforcement learning: MDP formalization, core concepts & policy-gradient methods
- RL for LLM training: RLHF procedure
  - training and constructing reward models
  - RLAIF
- architecture of fine-tuned LLMs
- understanding LLMs:
  - construction of test sets
  - I/O evaluation on benchmarks
  - mechanistic interpretability
  - evaluation of uncertainty representation
  - RL component evaluation



Posters and projects

Examples

Groups of 3-5

Example 6 ECTS project:

- systematically investigate the effects of RLHF on linguistic performance
- run evaluations on syntax / semantics / pragmatics / reasoning benchmarks
  - base model
  - fine-tuned model
- compare results
  - e.g., statistical tests comparing accuracy

Example 9 ECTS project:

- understand how helpfulness is represented in RLHF set up
- analyse helpfulness of RM dataset
  - extract some examples where you judge that preferred answer is more helpful
  - analyse lexical / syntactic / form aspects
- analyse RM performance
- evaluate LM fine-tuned with this RM
  - likelihood of using preferred lexical / syntactic / formal aspects
  - look at different tasks

More info during the break!

Merry Christmas!