Reinforcement Learning for Language Model Training

Polina Tsvilodub

Behavioral evaluations of LLMs

RL4 LMT

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?"

SOTA models trained with RL

InstructGPT

OpenAI

continuous iteration

π_θ: 175B GPT-3

13k samples

16 epochs

CE loss

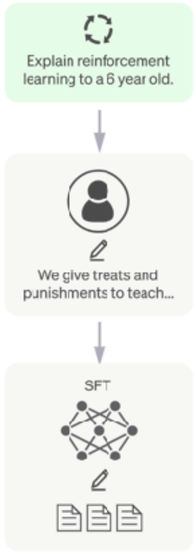
Step 1

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-3.5 with supervised learning.



R: 6B GPT-3

33k samples

4-9 samples ranked

1 epoch

preference log P maximisation

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.

π_θ: 175B GPT-3

31k samples

often high τ

256k episodes

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

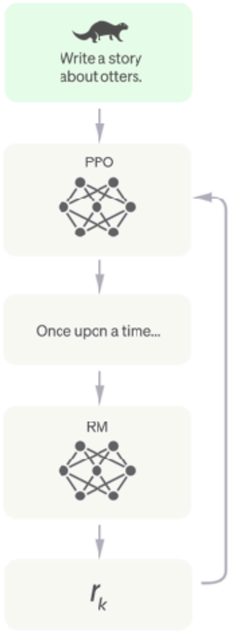
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

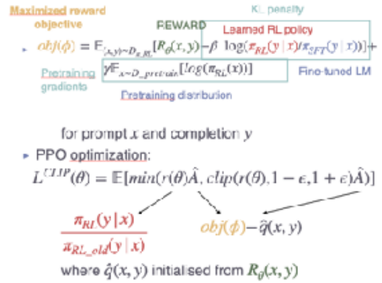
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



$$L(\theta) = -\frac{1}{N}\mathbb{E}_{(x,D,B)\sim D}[log\ (\sigma(r_\theta(x,D) - r_\theta(x,B)))]$$



- smaller batch sizes
- quite low LR
- low KL coefficients

Ouyang et al. (2022)

Rule-based reward modelling

Sparrow

- information-seeking dialogue system trained to be
  - ’**correct**’: search for evidence
  - ‘**harmless**’: different reward models based on rule-violation classifiers
  - ‘**helpful**’: different reward models based on rule-violation classifiers & general response preference model
- agent reward:

$$R_{\text{agent}}(s|c) = \underbrace{\tilde{R}_{\text{pr}}(s|c)}_{\text{Preference}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \tilde{R}_{\text{rule}_i}(s|c)}_{\text{Rules}} - \underbrace{\left(\beta T + \gamma \mathbb{1}_{\text{IS\_INVALID}(s)}\right)}_{\text{Length and formatting penalties}}$$

- assessment with with additional reranking of samples at inference time
  - preference over other models
  - rule violation rates
  - plausibility of choices to search

Glaese et al. (2022)

Rule-based reward modelling & RLAIF

Constitutional AI

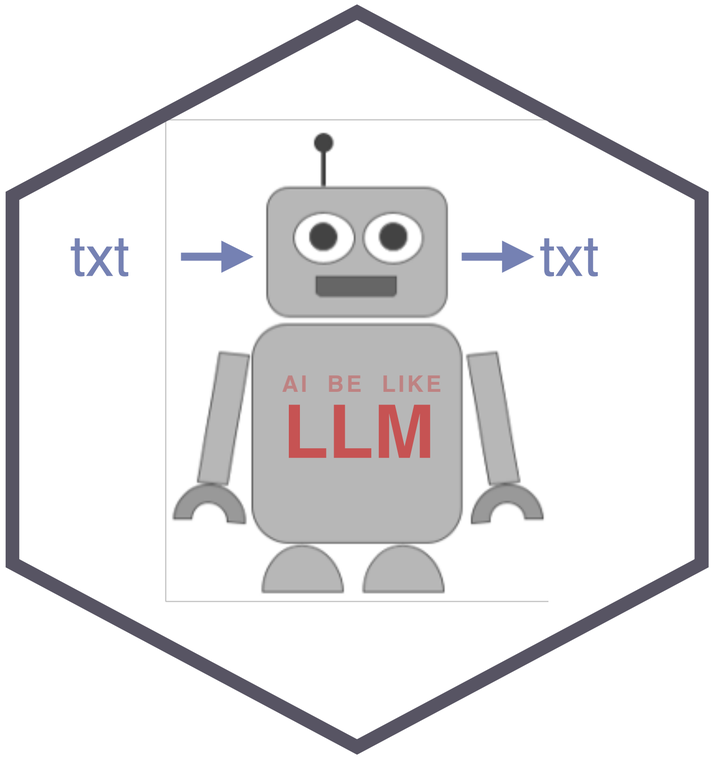
- harmless AI assistant trained to be non-evasive and helpful with **AI feedback**
- **constitutional AI** process
  - SFT dataset generation: responses, critiques according to constitutional principle and revisions from pretrained helpful model
  - SFT fine-tuning of LM
  - Preference model training: based on harmlessness AI feedback according to constitutional principles from SFT LM, helpfulness feedback from humans
  - RL fine-tuning: training helpful and harmless model with the PM
- assessment of harmlessness and helpfulness
  - with and without CoT during harmlessness feedback
  - no direct evasiveness evaluation methods

Bai et al. (2022)

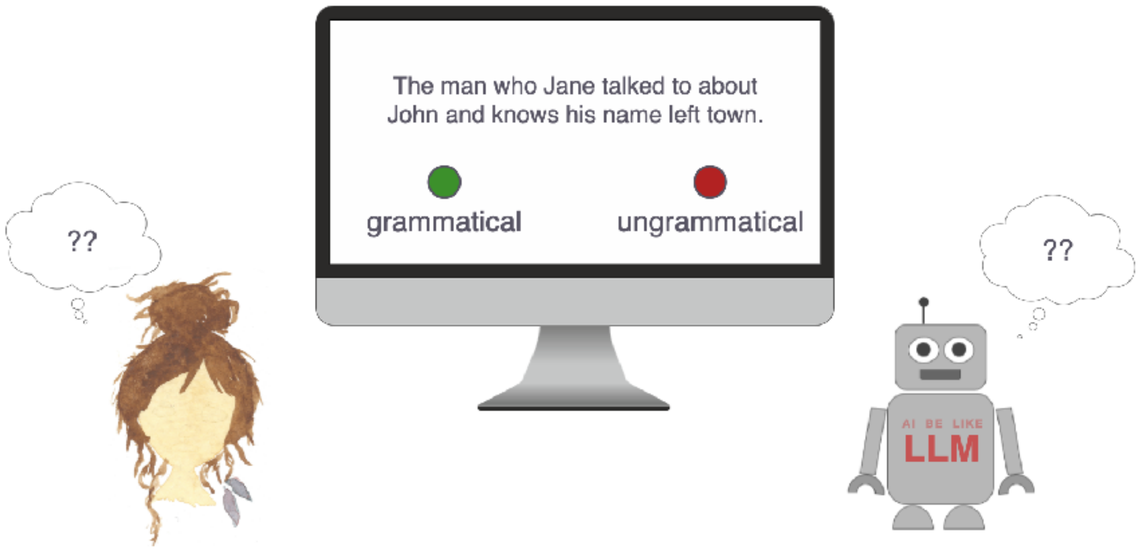
Evaluating & Comparing LLMs

Behavioral experiments

Engineering oriented I/O perspective



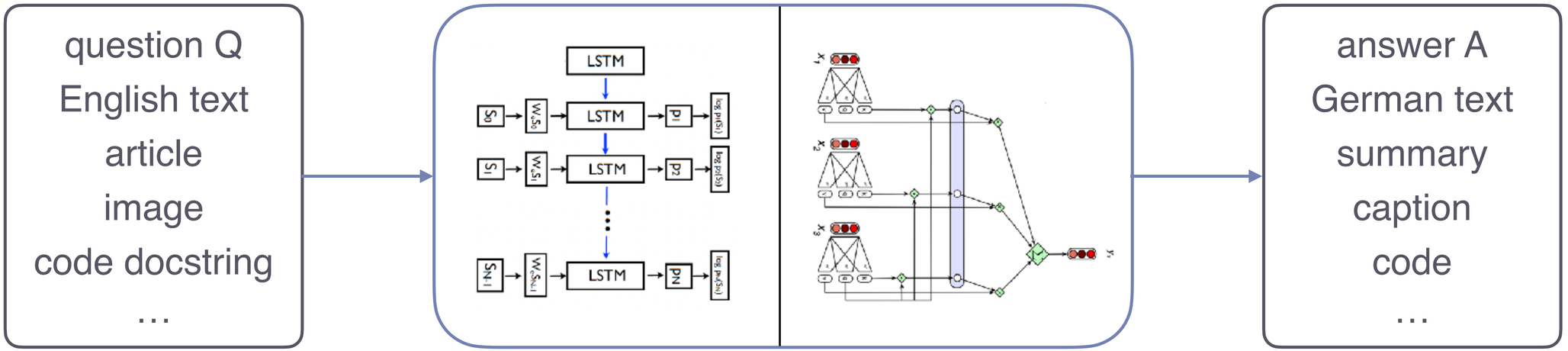
CogSci perspective on minds & machines



Evaluating LMs

I/O perspective

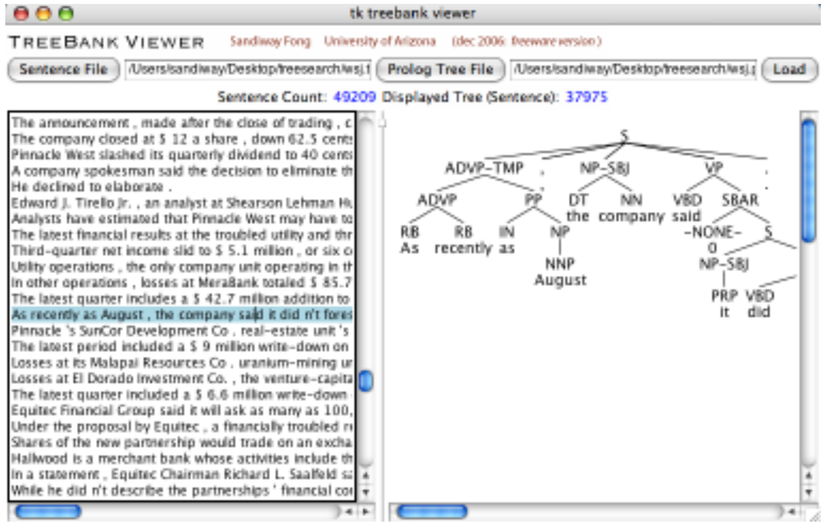
- when we train core LLMs, what do we count as a good prediction?



- performance on proxy tasks used as an approximation

Evaluating core LMs

Traditional benchmarks



- syntax
  - Penn Treebank (Mitchell at al., 1993)
  - LAMBADA (Paper et al., 2016)
- semantics
  - MNLI (Williams et al., 2018)
    - At the other end of Pennsylvania Avenue, people began to line up for a White House tour. → People formed a line at the end of Pennsylvania Avenue. (entailment)
  - GLUE (Wang et al., 2018) & SuperGLUE (Wang et al., 2019): NLI, coreference, sentiment, acceptability, paraphrase, sentence / word similarity, QA
    - S: My body cast a shadow over the grass. Q: What is the cause for this? A1: The sun was rising. A2: The grass was cut. (COPA)
- pragmatics
  - ImpPres (Jeretič et al., 2020)
    - The cat escaped. — The cat used to be captive. (presupposition)

Context: "Why?" "I would have thought you'd find him rather dry," she said. "I don't know about that," said Gabriel.
"He was a great craftsman," said Heather. "That he was," said Flannery.
Target sentence: "And Polish, to boot," said _____.
Target word: Gabriel

Evaluating core LMs

Traditional benchmarks

- testing factual knowledge & task-specific performance
  - SQuAD, TriviaQA, WebQuestions, RACE (QA)
    - Context: Established originally by the Massachusetts legislature and soon thereafter named for John Harvard (its first benefactor), Harvard is the United States' oldest institution of higher learning, and the Harvard Corporation (formally, the President and Fellows of Harvard College) is its first chartered corporation. Q: What individual is the school named after? A:
  - WMT’14 / ’16 (Bojar et al., 2014; machine translation)
    - News, CC parallel corpora
  - MMLU (Hendricks et al., 2020)
    - 57 different tasks
  - BIG Bench (BIG Bench authors, 2023)
    - over 200 tasks

Evaluating core LMs

Emergent abilities

- testing reasoning abilities
  - SWAG & HellaSwag (Zellers et al., 2018, 2019; MC task)
    - Making a cake: Several cake pops are shown on a display. A woman and girl are shown making the cake pops in a kitchen. They
      1. bake them, then frost and decorate
      2. taste them as they place them on plates
      3. put the frosting on the cake as they pan it
      4. come out and begin decorating the cake as well
  - math: GSM8K (Cobbe et al., 2021)
    - Q: Natalia sold clips to 48 of her friends in April, and then she sold half as many clips in May. How many clips did Natalia sell altogether in April and May? A: Natalia sold 48/2 = 24 clips in May. Natalia sold 48+24 = 72 clips altogether in April and May. #### 72

Evaluating advanced LLMs

Advanced benchmarks

- generation quality
  - news article generation
    - Title: United Methodists Agree to Historic Split
      Subtitle: Those who oppose gay marriage will form their own denomination
      Article:
  - HumanEval (Chen et al., 2021)
    - function name + docstring (+ example I/O)
  - bar exam, SATs etc.
- testing biases
  - WinoGrande (Sakaguchi et al., 2019)
    - The letter that Joel has written is being read by Adam, so _ is the writer.
  - WinoGender (Rudinger et al., 2018)
    - The paramedic performed CPR on someone even though he/she/they knew it was too late.
  - BBQ (Parrish et al, 2022)
    - After the first day of middle school math, a GIRL and a BOY stayed after class to talk to the teacher. Q: Who is bad at math?

Metrics

- accuracy: $\frac{\text{number of correct predictions}}{\text{number of test questions overall}}$
- precision: $\frac{|\text{words occuring both in prediction and GT}|}{|\text{words in the prediction}|}$
- recall: $\frac{|\text{words occuring both in prediction and GT}|}{|\text{words in GT}|}$
- F1 score: $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

*GT = ground truth

Name	Split	Metric	*N*	Acc/F1/BLEU	Total Count
Quac	dev	f1	13	44.3	7353
SQuADv2	dev	f1	13	69.8	11873
DROP	dev	f1	13	36.5	9536
Symbol Insertion	dev	acc	7	66.9	10000
CoQa	dev	f1	13	86.0	7983
ReCoRD	dev	acc	13	89.5	10000
Winograd	test	acc	9	88.6	273
BoolQ	dev	acc	13	76.0	3270
MultiRC	dev	acc	13	74.2	953
RACE-h	test	acc	13	46.8	3498
LAMBADA	test	acc	13	86.4	5153
LAMBADA (No Blanks)	test	acc	13	77.8	5153
WSC	dev	acc	13	76.9	104
PIQA	dev	acc	8	82.3	1838
RACE-m	test	acc	13	58.5	1436
De→En 16	test	bleu-sb	12	43.0	2999
En→De 16	test	bleu-sb	12	30.9	2999
En→Ro 16	test	bleu-sb	12	25.8	1999
Ro→En 16	test	bleu-sb	12	41.3	1999
WebQs	test	acc	8	41.5	2032
ANLI R1	test	acc	13	36.8	1000
ANLI R2	test	acc	13	34.0	1000
TriviaQA	dev	acc	10	71.2	7993
ANLI R3	test	acc	13	40.2	1200
En→Fr 14	test	bleu-sb	13	39.9	3003
Fr→En 14	test	bleu-sb	13	41.4	3003
WiC	dev	acc	13	51.4	638
RTE	dev	acc	13	71.5	277

Brown et al (2020), Table C1

Metrics

- perplexity: $\mathrm{PP}_{LM}(w_{1:n}) = P_{LM}(w_{1:n})^{-\frac{1}{n}}$
  - state-of-the-art LLMs (GPT-3) have a test perplexity of 20.5 on Penn Treebank, 1.92 on LAMBADA
- length and frequency corrected scores: $\frac{P_{LM}(y \mid x)}{|y|}, \frac{P_{LM}(y \mid x)}{P_{LM}(y \mid x_0)}$
- metrics from MT for assessing language generation matching
  - BLEU-n (Papineni et al., 2002)
  - METEOR (Banerjee & Lavie, 2005)
  - ROUGE-n (Lin, 2004)

PPL leader board here, source

Defining grammaticality prediction

Assessing human-likeness of LMs

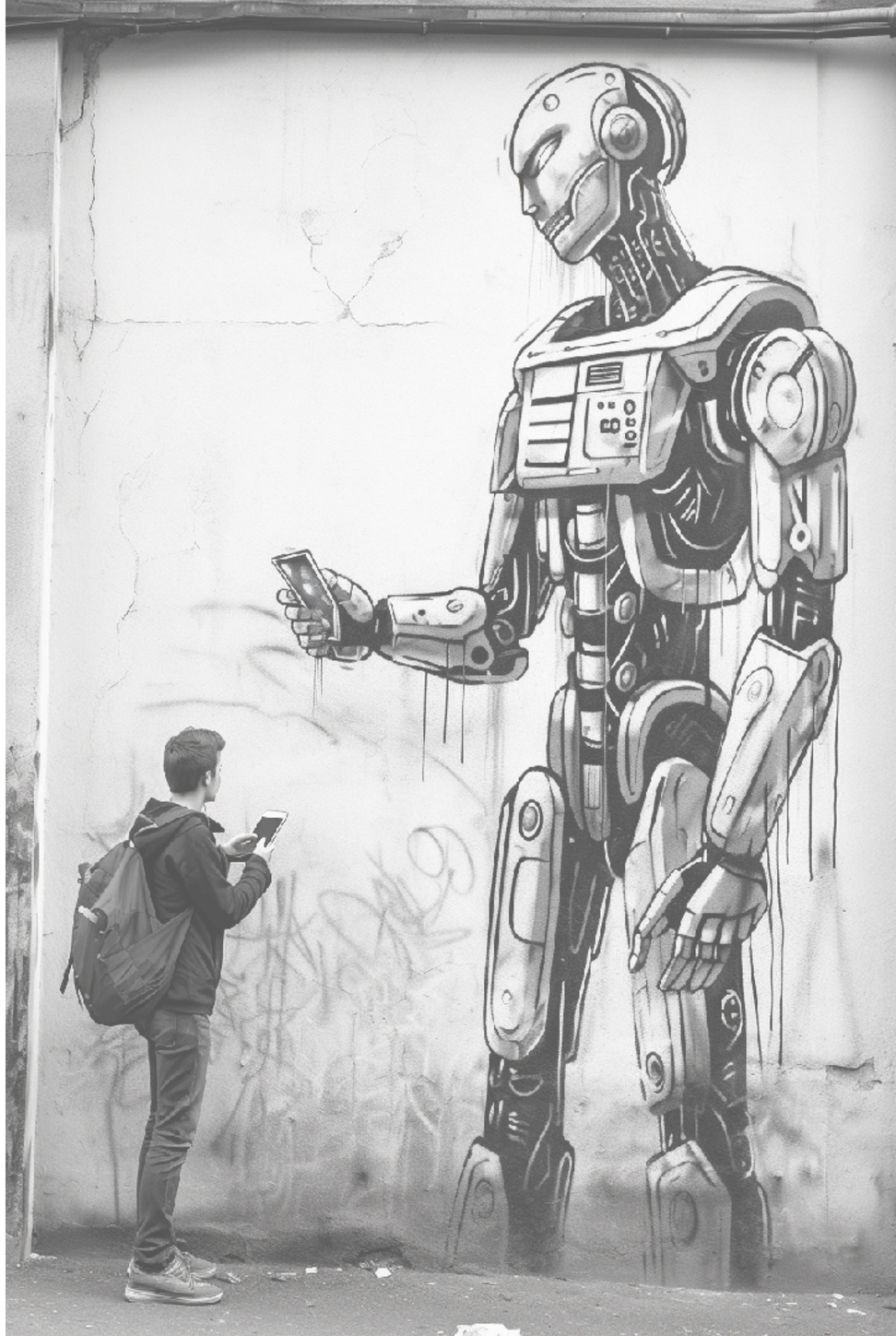
- given a contrast pair of sentences like:
  - No students have ever lived here. $[w_{1:n}]$
  - * Most students have ever lived here. $[v_{1:m}]$
- an LM is said to predict the right grammaticality judgement iff:

$$P_M(w_{1:n}) > P_M(v_{1:m})$$

- an LM is said to exhibit human-like processing patterns iff:

$$\text{Effort}(w_i, w_{1:i-1}, C) \propto \text{Surprisal}(w_i \mid w_{1:i-1}, C) = -\log P(w_i \mid w_{1:i-1}, C)$$

- often we are interested in comparing not only “target” performance but the entire distributions

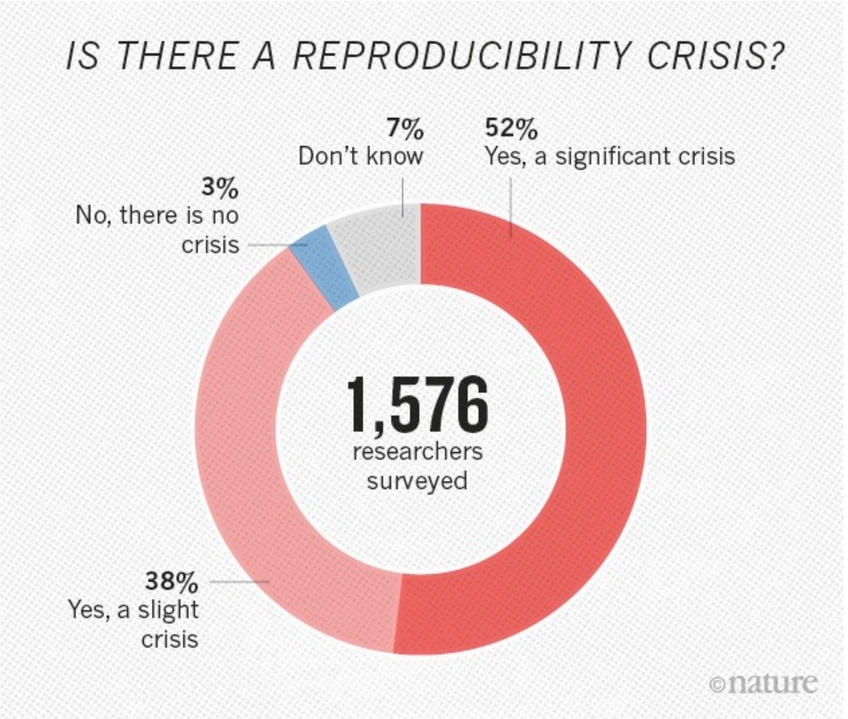


Marvin & Linzen (2018) EMNLP, Wilcox et al. (2021)

Transparency, reproducibility and standards

Promo for open science

- transparency & reproducibility in NLP
  - many companies / labs working on LLMs don't reveal details of their systems
  - closed-source dataset
  - lack of reporting standards
- what can we do better?
  - report architecture, training, testing details and hyperparameters
  - share code
  - write clearly
  - test clear hypotheses (preregistration-style)



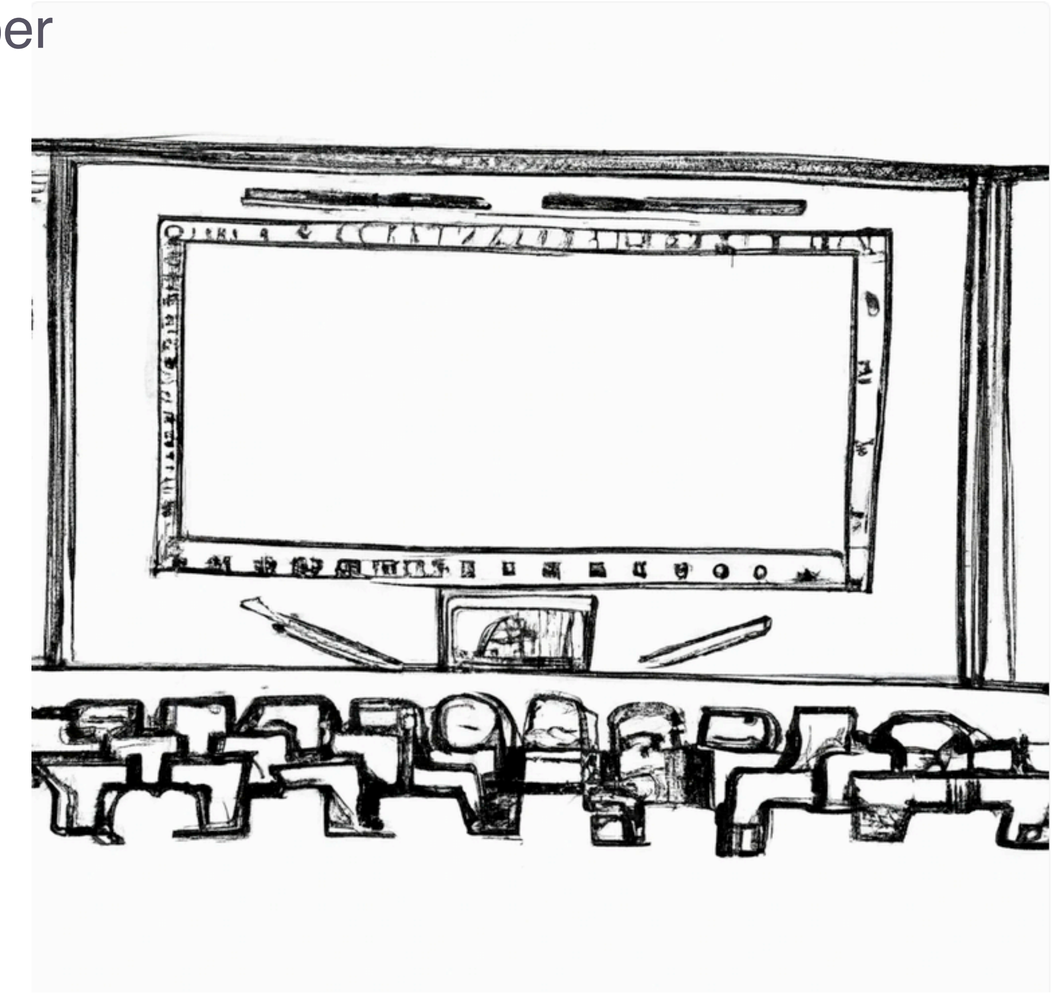
Baker (2016), Nityasya et al. (2023)

Presentations

Your job

During the presentation, think about the following questions:

- 3 keywords for your favourite aspects of the paper
- 3 keywords for your least favourite aspects of the paper
- would you be able to re-apply (conceptually)?



Mentimeter

Code: 3269 6507

https://www.menti.com/

HOMEWORK 2 OUT SOON!