Understanding Large Language Models

Carsten Eickhoff, Michael Franke and Polina Tsvilodub

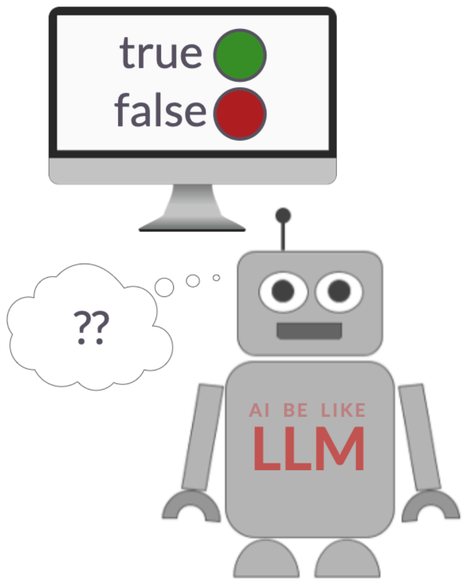
Session 08: Behavioral assessment

ie. study systematically the input & output behaviours of LLMs,
to understanding what are the doing in the behavioural level.

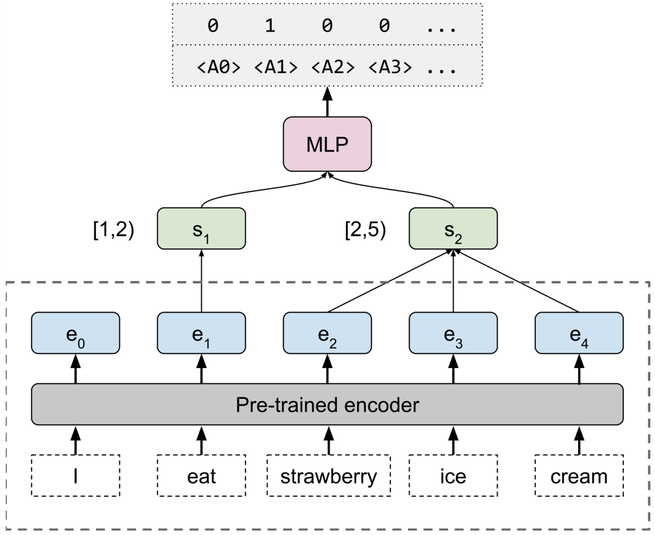
the internal representation.

Assessing what LLMs “know” and “can do”

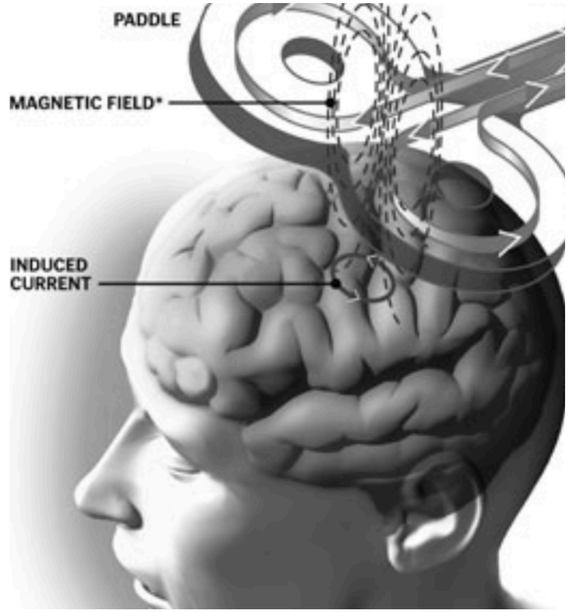
Behavior. Tests



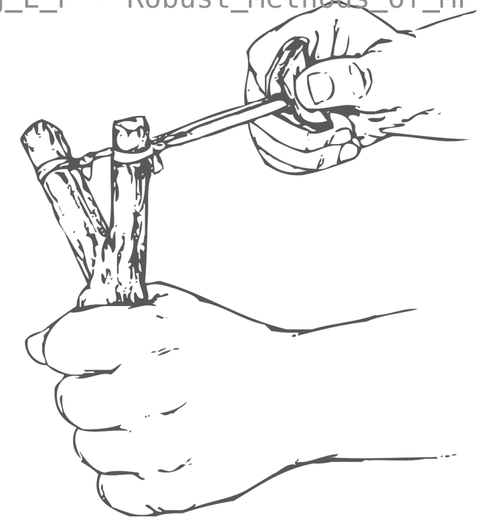
Probing



Intervention



Main learning goals



1. benchmark testing
2. evaluation metrics — accuracy scoring, calibration, etc
3. machine psychology
4. machine psycholinguistics — a special kind of machine psychology
5. reflection on methods

Behavioral assessment: Motivation

Assessing the behavior of a trained LM

2 ways to look at the i/o behaviours:

- is the I/O behavior in particular **applications**:
  - correct
  - trustworthy
  - reliable
  - robust doesn't break in in reasonable edge cases
  - aligned with human expectation / values
  - … important to note that the standard of these could vary from applicatoins
- is the I/O behavior **human like**:
  - correct when humans are correct
  - incorrect when humans are incorrect
  - of the same type as what humans do
  - …

Behavioral assessment: Motivation

Assessing the behavior of a trained LM

- is the I/O behavior in particular **applications**:
  - correct
  - trustworthy
  - reliable
  - robust
  - aligned with human expectation / values
  - …

⇒

- good applications
  - correct
  - safe
  - …

- is the I/O behavior **human like**:
  - correct when humans are correct
  - incorrect when humans are incorrect
  - of the same type as what humans do
  - …

⇒

- when/ how to anthropomorphize?
- understand how system works
- shed light on human by comparison to machine cognition

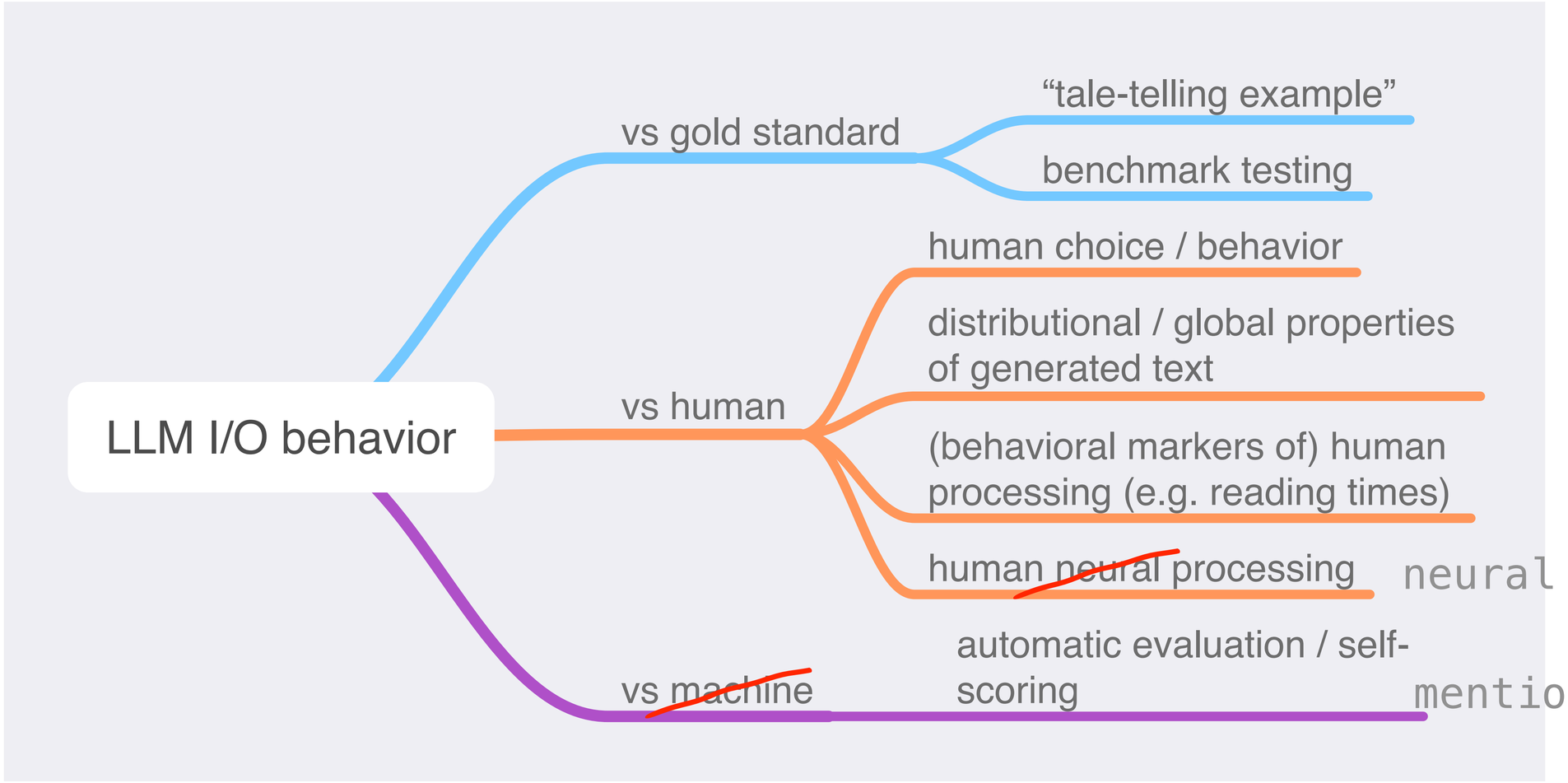
Common sense “axiom” of behavioral assessment

for downstream applications

*Downstream applications是指在完成初步模型训练或预处理后，使用该模型在特定实际任务中的应用，例如情感分析、机器翻译、文本生成等。

If we want to assess the behavior of system S because we want to know how well S performs in an application context C, then we should test the behavior of S under the exact same circumstances as S would be used in the application context C.

Overview


LLM I/O behavior
vs gold standard
“tale-telling example”
benchmark testing
vs human
human choice / behavior
distributional / global properties of generated text
(behavioral markers of) human processing (e.g. reading times)
human neural processing
neural markers
vs machine
automatic evaluation / self-scoring
mentioned only

“Tale-telling” by example

Motivating examples

in research papers

by different individuals. LMs additionally exhibit some success in relating beliefs to plans:

> ***Lou leaves a lunchbox in the work freezer. Syd, who forgot to pack lunch, eats Lou's instead. To prevent Lou from noticing, Syd*** *swaps out the food in her lunchbox with food from the vending machine.*

This completion that describes multiple individuals' differing motivations and beliefs along with information about the environment in which they act.

What these examples suggest, and what I want to argue in the rest of this paper, is that LMs can serve as **models of agents** in a narrow sense: they can predict relations between agents' observations, internal states, and actions or utterances. In partic-

single prompt completion

interpretation

serves as motivation for more in-depth analysis

Andreas (2022) "Language Models as Agent Models"

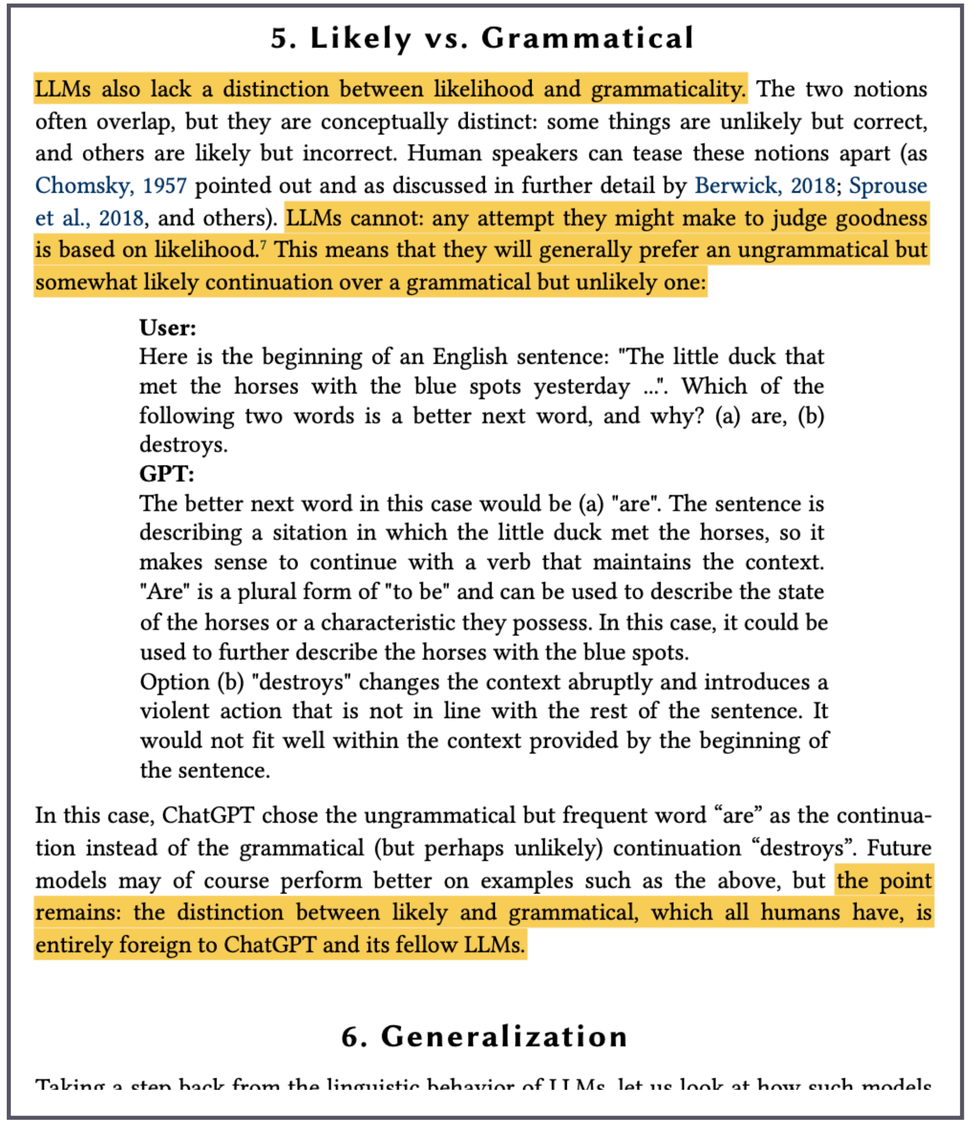
Critical failures (e.g., ethical considerations)

on social media

from here

Single examples as key arguments

in scientific writing



5. Likely vs. Grammatical

LLMs also lack a distinction between likelihood and grammaticality. The two notions often overlap, but they are conceptually distinct: some things are unlikely but correct, and others are likely but incorrect. Human speakers can tease these notions apart (as Chomsky, 1957 pointed out and as discussed in further detail by Berwick, 2018; Sprouse et al., 2018, and others). LLMs cannot: any attempt they might make to judge goodness is based on likelihood.[7] This means that they will generally prefer an ungrammatical but somewhat likely continuation over a grammatical but unlikely one:

> **User:**
> Here is the beginning of an English sentence: "The little duck that met the horses with the blue spots yesterday ...". Which of the following two words is a better next word, and why? (a) are, (b) destroys.
> **GPT:**
> The better next word in this case would be (a) "are". The sentence is describing a sitation in which the little duck met the horses, so it makes sense to continue with a verb that maintains the context. "Are" is a plural form of "to be" and can be used to describe the state of the horses or a characteristic they possess. In this case, it could be used to further describe the horses with the blue spots.
> Option (b) "destroys" changes the context abruptly and introduces a violent action that is not in line with the rest of the sentence. It would not fit well within the context provided by the beginning of the sentence.

In this case, ChatGPT chose the ungrammatical but frequent word “are” as the continuation instead of the grammatical (but perhaps unlikely) continuation “destroys”. Future models may of course perform better on examples such as the above, but the point remains: the distinction between likely and grammatical, which all humans have, is entirely foreign to ChatGPT and its fellow LLMs.

6. Generalization

Taking a step back from the linguistic behavior of LLMs, let us look at how such models

very general issue

single example from a single model with a single method of assessment

strong conclusion

Katzir (2023)

Taking stock: “tale-telling” by example

- good for motivation or suggestive demonstration
- strong conclusions only if
  - a single example is evidence enough
    - e.g., to refute a universal statement
    - e.g., system must not get a single instance wrong *ever*
  - back-up by more systematic testing
    - several items
    - several models
    - different assessment methods
    - ideally: **antagonistic testing**

Benchmarks

Benchmarks for I/O testing

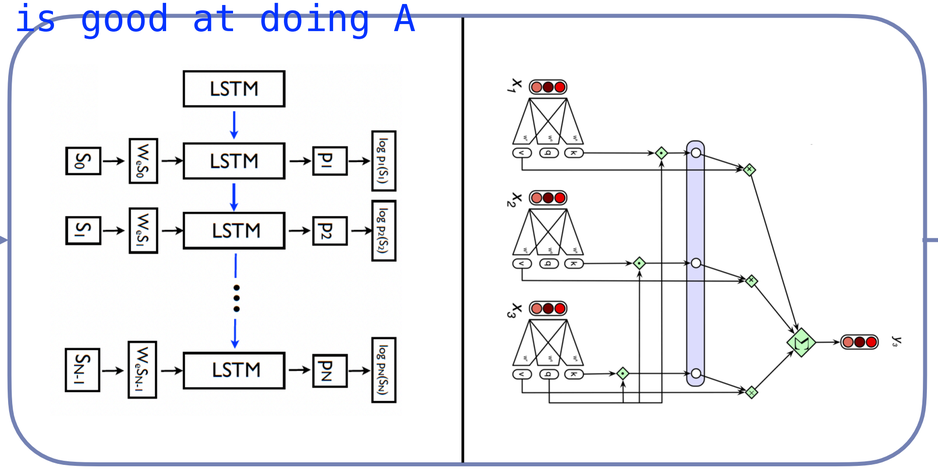
- what is it?
  - standardized **set of tasks** or datasets with fixed **standards of evaluation**
  - designed to evaluate performance & capabilities

what we want to do under I/O testing is how (under systematic variations of) the input is ideally designed to show aspects from the output.
How different system produces outputs.
Provide some kind of measurement for the input-output pairs in order to draw some quantative conclustions on the behaviours of a single system to compare various systems. So able to say the system is good at doing A but not so well on B.

- what is it **for**?
  - systematic assessment of (emergent) capability
  - streamlined comparison of different systems
    - e.g., **scaling** (performance boost from size *ceteris paribus*)
  - track advancement in the field
  - identify strengths and weaknesses
  - feedback for future work
  - define "what the community cares about"

其他条件相同

question Q
English text
article
image
code docstring
…



answer A
German text
summary
caption
code
…

input **output**

Types of common benchmarks

`list of topics of which there are benchmarks`

- **Common sense reasoning**

  BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018).

- **Question answering (closed-book)**

  Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017)

- **Natural language understanding**

  RACE (Lai et al., 2017), SQuAD (Rajpukar et al. 2016) …

- **Mathematical reasoning**

  MATH (Hendrycks et al., 2021) and GSM8k (Cobbe et al., 2021) …

- **Code generation**

  HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021)

- **Linguistic abilities**

  ImpPres (Jeretic et al. 2020), CommitmentBank (de Marneffe et al. 2019), BLiMP (Warstadt et al. 2020), …

- **Domain-specific expert (world) knowledge**

  MMLU (Hendrycks et al., 2020) …

Common benchmark collections

older

- **GLUE** general language understanding evaluation
  - Wang et al. (2018, website, huggingface)
  - 9 tasks
    - CoLA, SST-2, MRPC, QQP, STS-B, MNLI, QNLI, RTE, WNLI

older

- **SuperGLUE** general language understanding evaluation
  - Wang et al. (2019, website, huggingface)
  - 8 tasks
    - BoolQ, CommitBank, COPA, MultiRC, ReCorRD, RTE, WiC, WiSC

the biggest and most salient set of tasks

- **BIG-Bench** beyond the imitation game
  - Srivastava et al. (2023, repo)
  - 204+ tasks (all sorts of topics)
- **HELM** holistic evaluation of language models
  - Srivastava et al. (2023, repo)
  - 51 tasks (HellaSwag, BoolQ, MMLU …)
  - **metrics**: accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency

BIG-Bench

task types and evaluation metrics

consists of ... the % could change by urself.

- JSON tasks (~80%) u specify the task in a json file
  - multiple-choice / classification
    - (weighted) accuracy
    - expected calibration error
    - BRIER scores
  - generation
    - BLEU
    - BLEURT
    - ROUGE
    - exact match
- programmatic tasks (~20%)
  - python code to directly interact with model
    - generation & log-probs

2 major types of tasks

Example: Presuppositions as NLI

part of BIG-Bench

task description

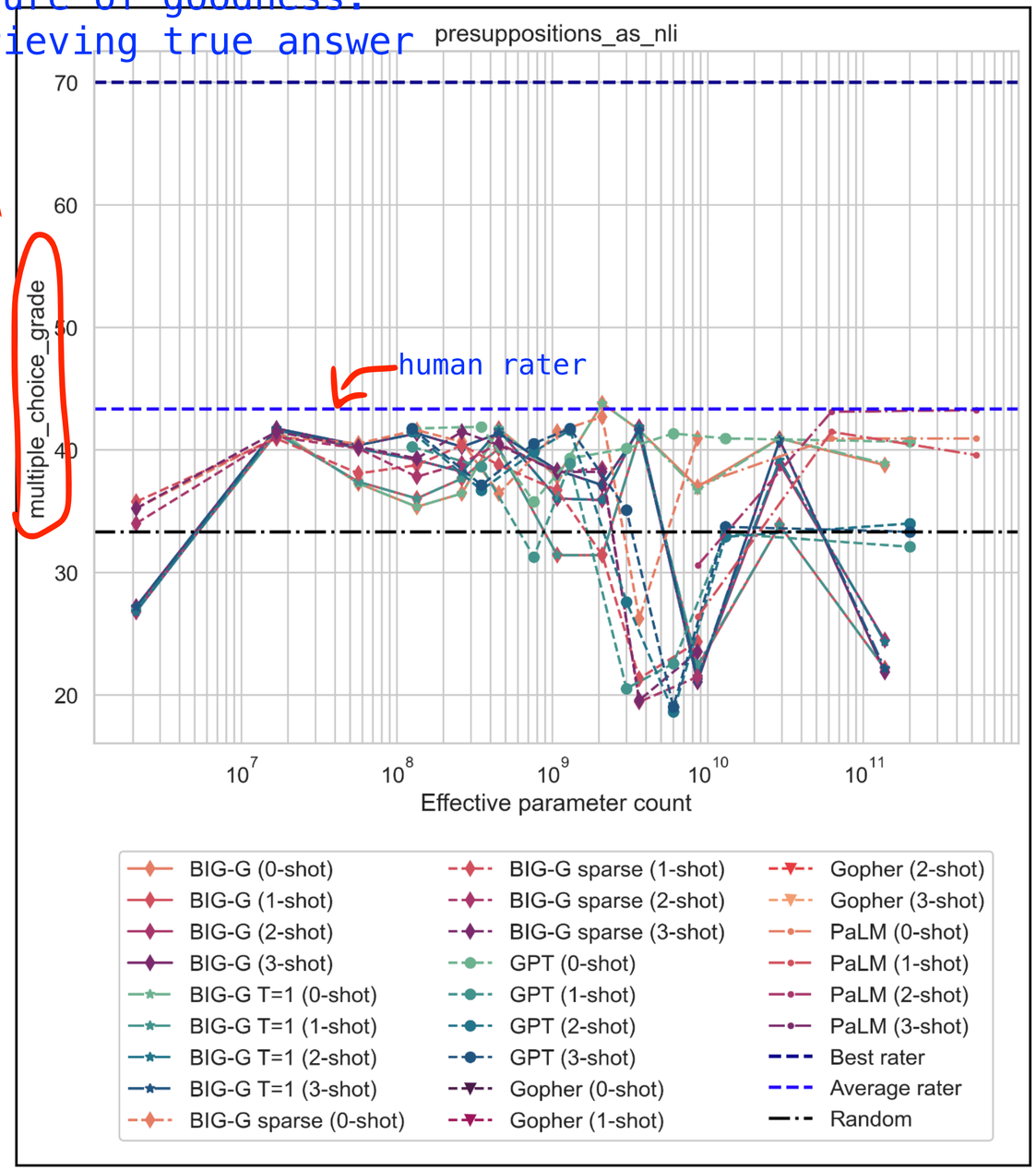
This is a natural language inference task. There are two sentences in English. The answer is "entailment" if the first sentence entails the second, "contradiction" if the second sentence contradicts the first, and "neutral" if neither is of those two cases holds.

Sentence 1: The cops had him in their headlights. He ran hard and fast, fiercely pumping his legs, his arms, but they gained on him quickly, swerving in front of him to block his way. Winded, aching, he didn't fall on his knees in the street.

Sentence 2: He was standing earlier.

(X) entailment ( ) contradiction ( ) neutral

essentially the measure of goodness:
accuracy of the retrieving true answer



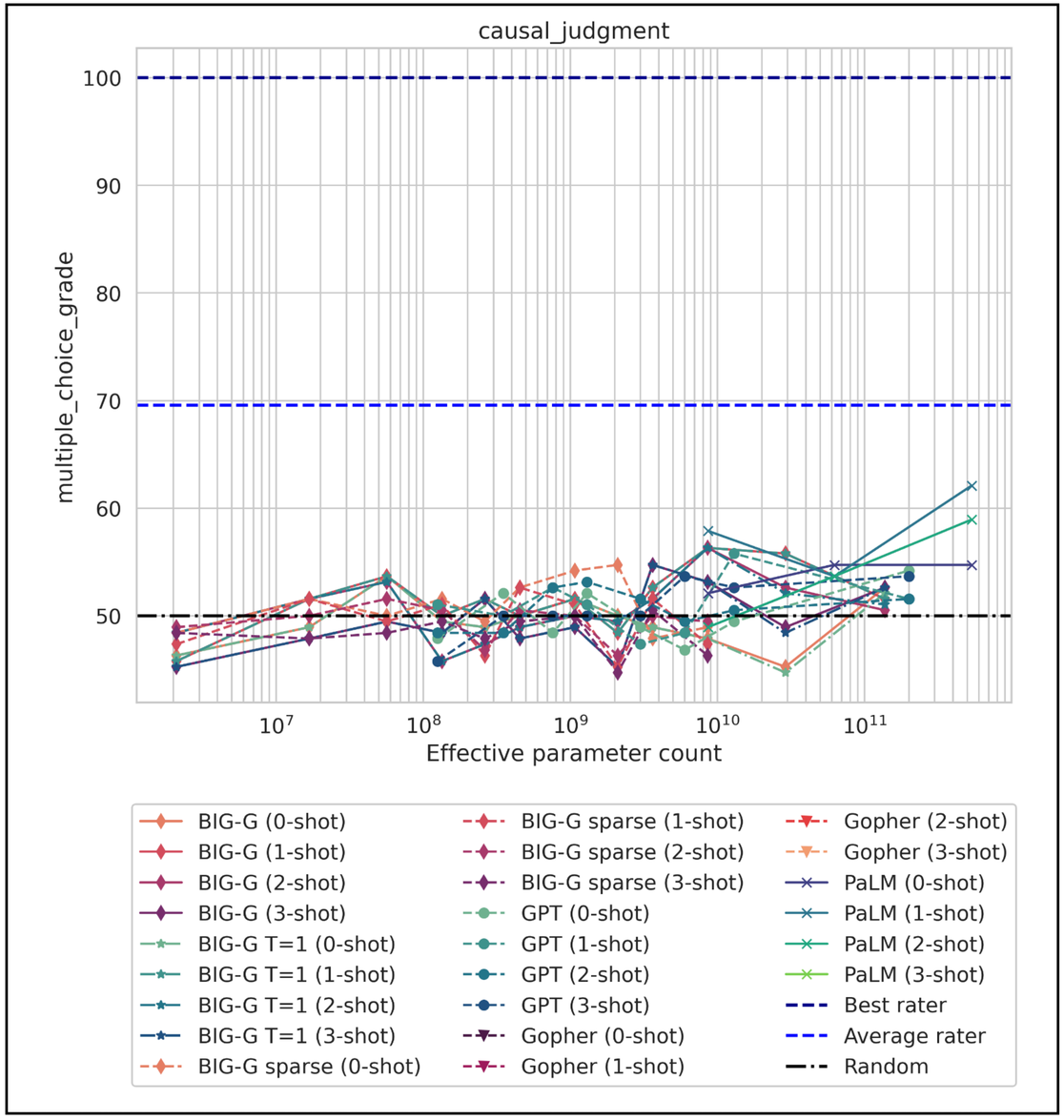
Example: Causal Judgement

part of BIG-Bench

How would a typical person answer each of the following questions about causation?

The CEO of a company is sitting in his office when his Vice President of R&D comes in and says, 'We are thinking of starting a new programme. It will help us increase profits, but it will also harm the environment.' The CEO responds that he doesn't care about harming the environment and just wants to make as much profit as possible. The programme is carried out, profits are made and the environment is harmed. Did the CEO intentionally harm the environment?

- (X) yes
- ( ) no



Example: Empirical Judgements

part of BIG-Bench

Determine whether a given sentence asserts a causal, correlative, or neutral relation between two events. If the sentence asserts a causal relation respond causal, if the sentence asserts a correlative relation respond correlative, if the sentence asserts neither a causal nor a correlative relation between two events respond neutral.

If the sun shines on the stone, the stone becomes warm.

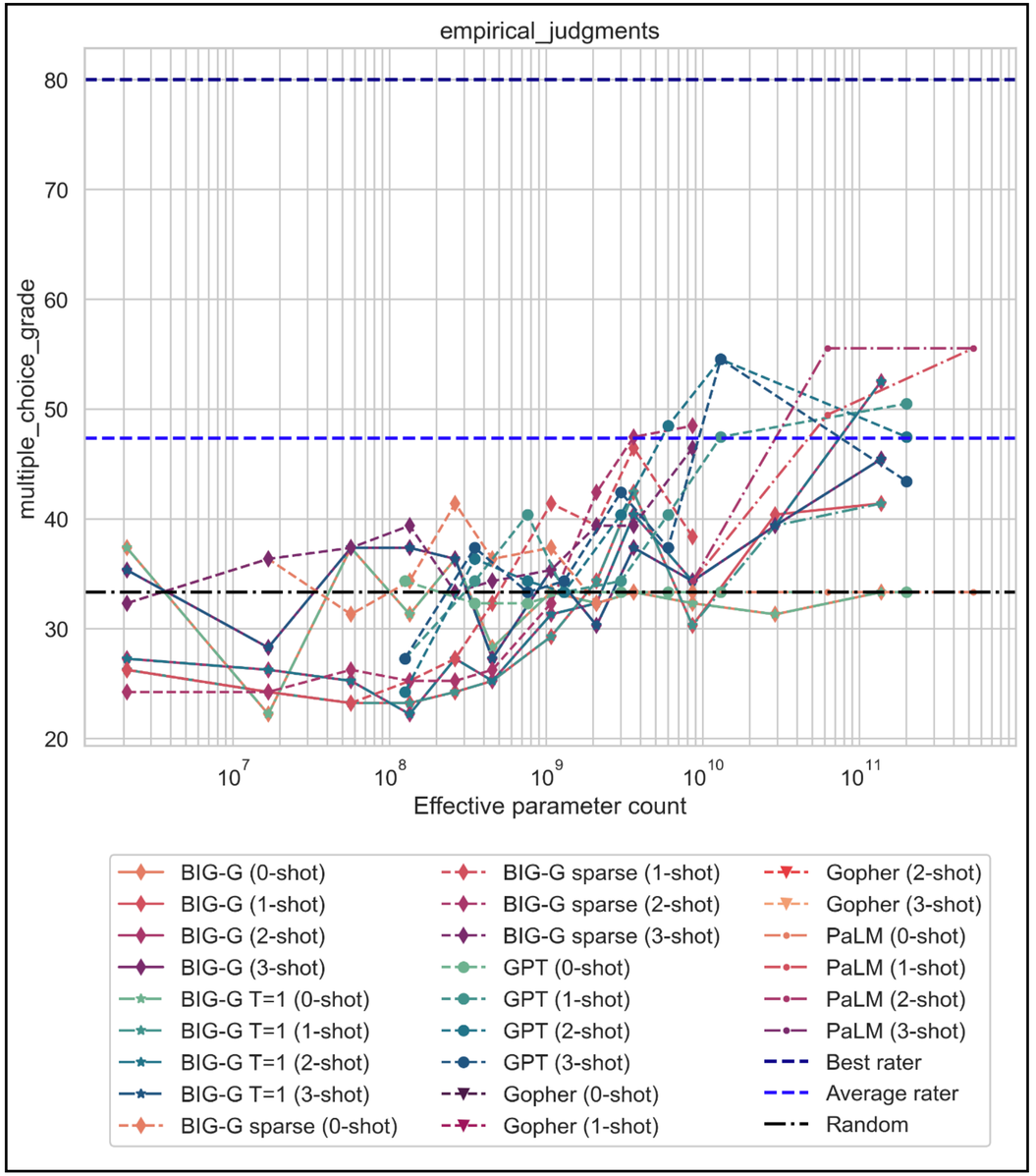
◯ causal ⓧ correlative ◯ neutral

The sun warms the stone.

ⓧ causal ◯ correlative ◯ neutral

If Boo is a cat, Boo is a feline.

◯ causal ◯ correlative ⓧ neutral



Example: ConLang translation

translate to adna

part of BIG-Bench

example of generation task

The following are sentences in Adna and their English translations:

ADNA: Ndengi ngase.

ENGLISH TRANSLATION: He drinks water.

ADNA: Ngoru ndatab ndengi ngase.

ENGLISH TRANSLATION: The child keeps drinking water.

...

Now, translate the following from English to Adna:

English: The teacher carries the water down.

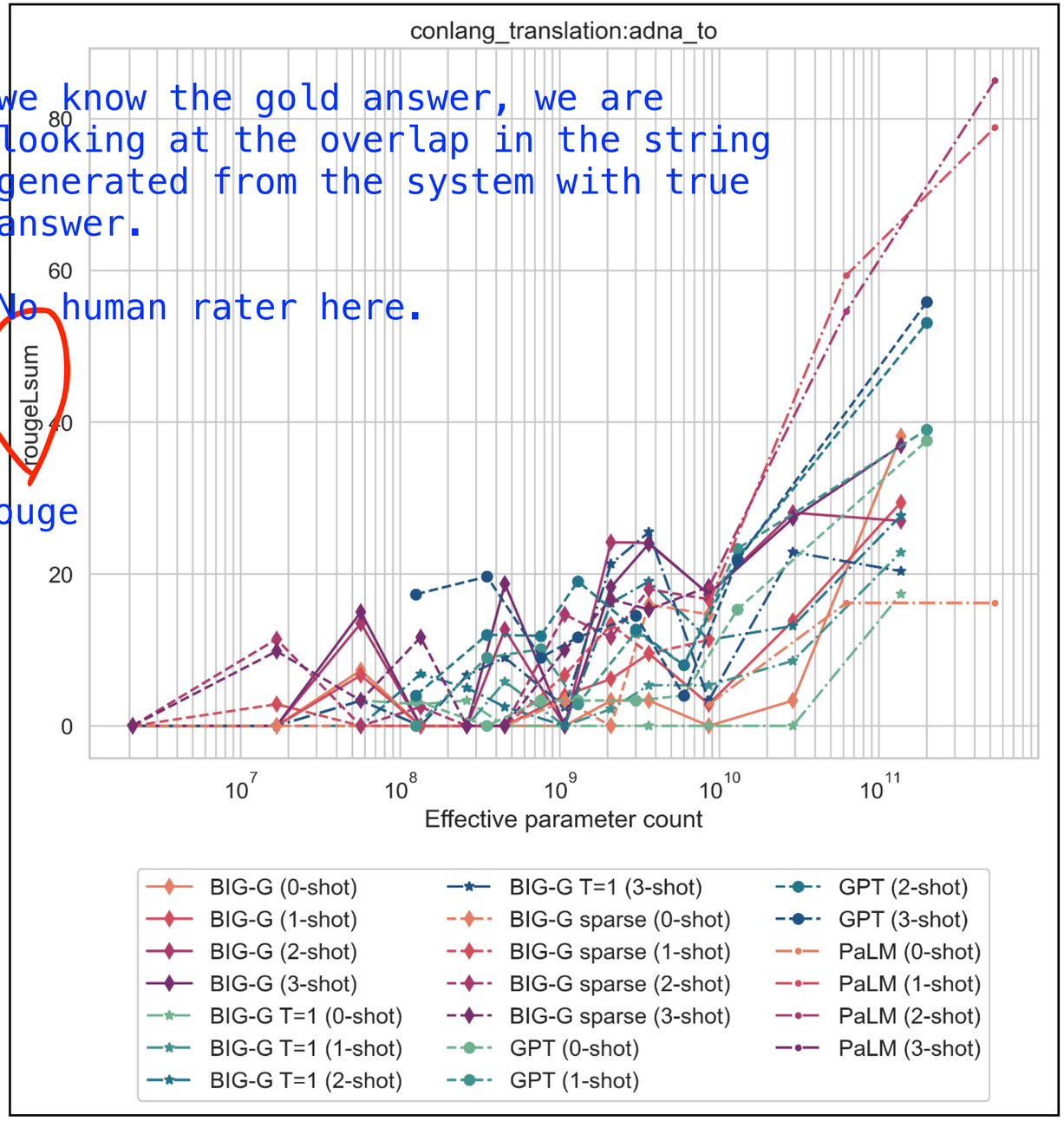
Adna:

[expected answer: Kansasres ndekaselib ngase ndisbo.]

we know the gold answer, we are looking at the overlap in the string generated from the system with true answer.

No human rater here.

Rouge



Taking stock: benchmark testing

- large scale assessment of system capabilities
- batteries of standardized tests
- systematics comparison (e.g., scale effects)
- sets ‘common research agenda’
- individual tests can be of poor-ish quality
  - benchmark test called “X” is not necessarily carefully designed to test for ability “X”

Evaluation metrics
when a gold-standard exists

Accuracy (standard winner-takes all version)

in multiple-choice / classification tasks

arg-max acc

- set of **items** I
- each **item** $i \in I$ consists of
  - x_i input prompt (string)
  - $y_{i1}, \ldots, y_{ik}$ set of answer options (strings)
  - designated "true" option t_i
- **raw score** for each option:

how probable is a particular answer

$$s(y_{ij}) = \sum_{l=0}^{|y_{ij}|} \log P\left(y_{ij} \mid x_i\right)$$

- choice of option via arg-max:

$$c_i = \arg\max_j s\left(y_{ij}\right)$$

then look at the one with the maximun score

- accuracy: $\mathbb{E}_i \; \delta_{c_i = t_i}$

input prompt x_i

This is a natural language inference task. There are two sentences in English. The answer is "entailment" if the first sentence entails the second, "contradiction" if the second sentence contradicts the first, and "neutral" if neither is of those two cases holds.

Sentence 1: The cops had him in their headlights. He ran hard and fast, fiercely pumping his legs, his arms, but they gained on him quickly, swerving in front of him to block his way. Winded, aching, he didn't fall on his knees in the street.

Sentence 2: He was standing earlier.

◯ entailment ◯ contradiction ◯ neutral

answer options y_{i1}, y_{i2}, y_{i3}

Accuracy (stochastic choice)
in multiple-choice / classification tasks

```
softmax acc
```

- **stochastic choice** via soft-max:

$$P(c_i = j) \propto \exp\left(\alpha \; s\left(y_{ij}\right)\right)$$

- accuracy (is now a function of α):

$$\mathbb{E}_i \; \mathbb{E}_{c_i} \; \delta_{c_i = t_i}$$

```
define accuracy as
expectation of all items
and expectation over choices u make
under ur probabilistic policy to be correct
```

Excursion: argMax- vs softMax-based accuracy

- two options: target & competitor
- target's score is a small ϵ better in 80% of the items
- in the remaining 20%, the competitor option has a virtually infinitely larger score
- argMax-accuracy is 80%
- softMax-accuracy is ~40%
  - in 80% of the cases target choice probability is ~.5, in the remaining 20% it is virtually 1
  - so: $0.8 \times 0.5 + 0.2 \times 0 = 0.4$

Confidence calibration 校准

an evaluation that big-bench and such use for their model

the idea:

- predicted category probabilities should approximate true probability

@GPT Confidence calibration是一种衡量机器学习模型预测可信度的方法，
目的是使模型预测的类别概率与实际的类别概率尽可能一致。
简单来说，就是希望模型给出的预测概率能真实反映其准确性。
例如，当模型给出某个类别有70%的概率时，我们希望在大量类似情况下，这个类别的出现频率确实接近70%。

calibration matrix

input data	erwtwge	pabyalkn	vpooiuy	vpooiuw
true label	A	A	B	B
prediction 1	0.6	0.51	0.49	wrong, A 0.51
prediction 2	0.6	0.51	0	wrong, A 1

Naturally, for similar input you have similar label

No need to sum up to 1. 0.49 for b being the correct label

model 1 & 2 makes the same mistable under argmax, but
model 2 is more over-confident,
model 1 is at least uncertain,
we would learn at least the uncertainty

Brier scores

- fix N items for a K-way categorization problem
- let p_{ij} be the model's predicted probability for item i belonging to category j
- one-hot encoding of true category:

$$d_{ij} = \begin{cases} 1 & \text{if category } j \\ 0 & \text{otherwise} \end{cases}$$

- **Brier score** for model's predictions:

$$\mathrm{BS} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \left(p_{ij} - o_{ij} \right)^2$$

input data	erwtwge	pabyalkn	vpooiuy	vpooiuw	**BS**
true label	A	A	B	B	
prediction 1	0.6	0.51	0.49	0.51	~0.505
prediction 2	0.6	0.51	0	1	~0.755

one way of measuring calibration

Expected calibration error

in multiple-choice / classification tasks

- set of multiple-choice items I
- true category for item $i \in I$ is t_i
- model's choice of item $i \in I$ is c_i
- model's probability for choice c_i is p_i
- expected calibration error (ECE) is defined as:

$$\mathbb{E}_{i \in I}\left[p_i - \mathbb{E}_{i' \in I}\left(t_{i'} = c_{i'} \mid p_i = p_{i'}\right)\right]$$

proportion of that equal p=p being correct

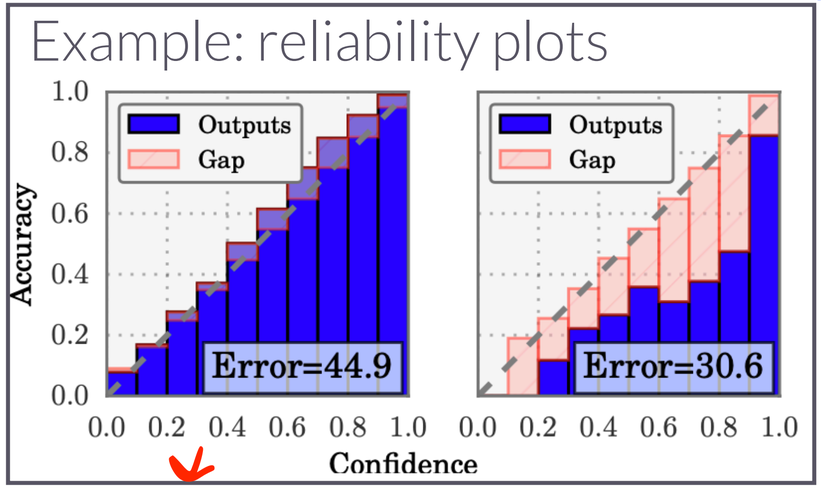
expectation over the whole dataset, for each item

expectation of that prediction actually being true

a correctness threshold/score of .6, look at all cases where model predicts .6

item	p_i	correct?	correct | p_i=0.6
1	0.3	❌	-
2	0.6	✅	✅
3	0.7	❌	-
4	0.6	❌	❌
5	0.2	✅	✅
6	0.6	✅	-
7	0.5	❌	-
8	0.6	✅	✅
…	…	…	…

Example: reliability plots



range of predictions from 0% to 20%

if calibrated, bins形状会接近该斜对角线，如左图，右图则是not calibrated.

$$\mathbb{E}_{i' \in I}\left(t_{i'} = c_{i'} \mid p_i = p_{i'}\right)$$

average correct

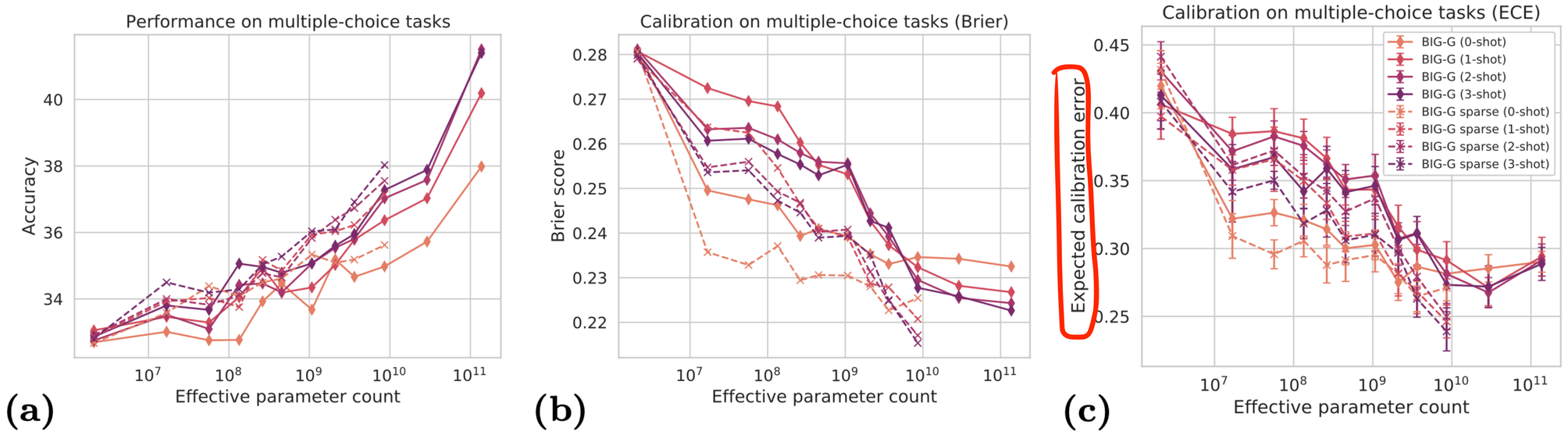
proportion of cases where the .6 prediction is correct. 此时p_i是.6.

Naeini et al. (2015), Guo et al. (2017)

Accuracy & calibration in LLMs

Note: even acc is high (based on argmax), they could still be over-confident about choices

2/ the expected calibration error is still rather high in absolute values



key takeaways: 1/

Figure 5: **Language models make poorly calibrated predictions, but calibration improves as the models are made larger.** **(a)** Performance is shown aggregated across all JSON multiple-choice tasks, along with the associated **(b)** Brier score and **(c)** expected calibration error (ECE) measures of model calibration. Error bars are obtained by 10-bootstrap sampling over evaluated tasks.

Srivastava et al. (2023, BIG-Bench)

Metrics for generation

- BLEU-n (Papineni et al., 2002)
  - co-occurence on n-grams between generated and reference sequences `have known preference towards shorter sentences`
- ROUGE-n (Lin, 2004)
  - similar to BLEU-n but on longest common sequence
- METEOR (Banerjee & Lavie, 2005)
  - harmonic mean of unigram precision and recall
  - matching target and output via exact matching, synonymy, stem-identity ...

- some weaknesses
  - depend on finite reference corpus
  - depend on tokeniser
  - might have biases towards particular form of candidate predictions
  - might not align well with human judgements

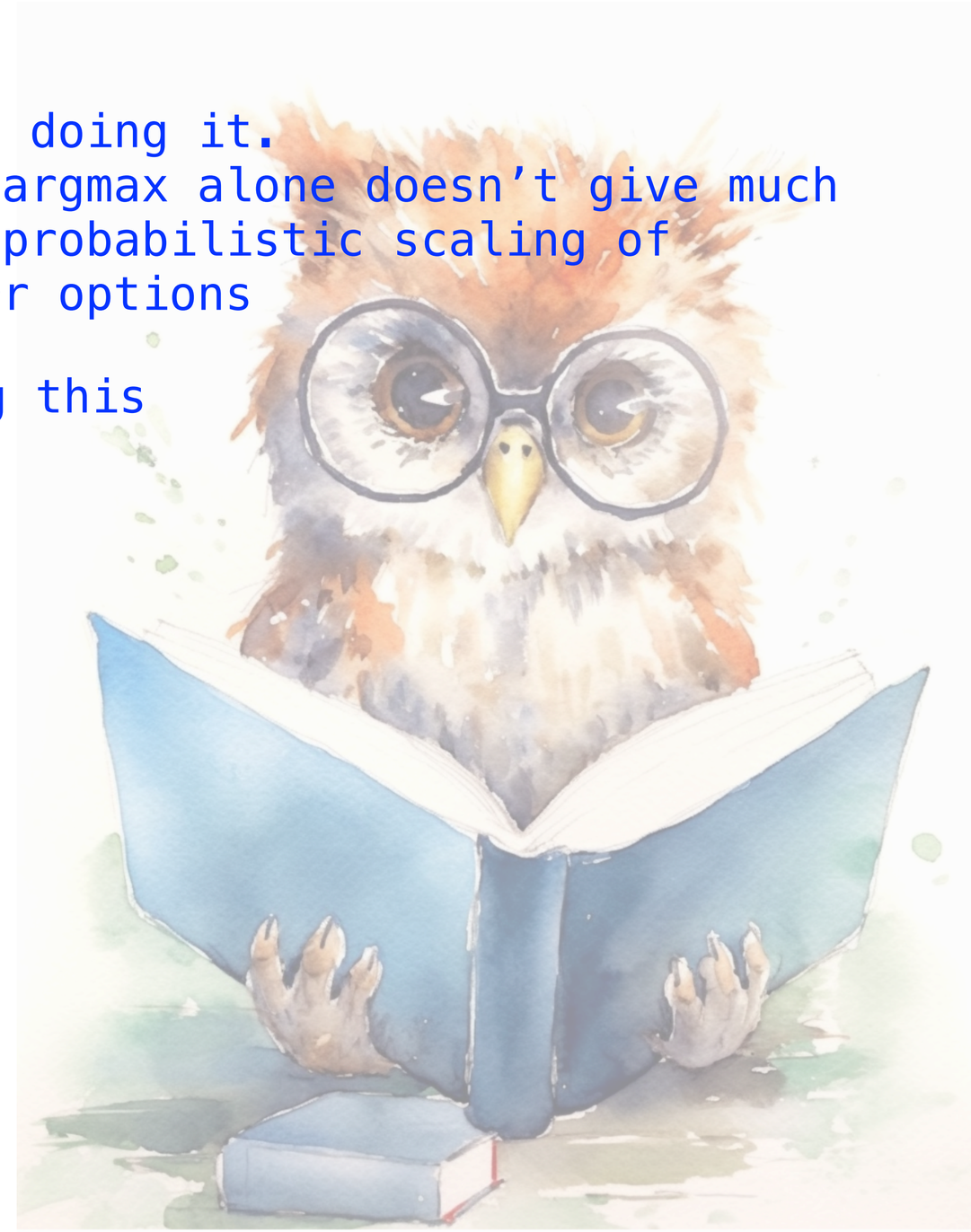
summary

Taking stock: common evaluation metrics

- accuracy of classification
  - based on argMax log-probabilities
- confidence calibration
  - Brier scores, expected calibration error
- quality of generations
  - BLEU, ROUGE etc.

standard way of doing it.
But looking at argmax alone doesn't give much info about the probabilistic scaling of different answer options

one way of addressing this



Note: acc and calib are correlated but conceptually different

Bias corrections

for multple choice classification

Length correction

- raw score for each option (previously):

$$s(y_{ij}) = \sum_{l=0}^{|y_{ij}|} \log P\left(y_{ij} \mid x_i\right)$$

log-probability of continuation

- length-corrected score:

$$s(y_{ij}) = \sum_{l=0}^{|y_{ij}|} \log P\left(y_{ij} \mid x_i\right)$$

typo!

perplexity of continuation

- length correction achieves better accuracy on common benchmarks
  - Brown et al. (2020, GPT-3)
- length correction is often default

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

- **perplexity**:

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

- **average surprisal**:

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

but this might not the only bias we have to correct for in our accuracy scoring… next page

Accuracy variance under spurious prompt variation

"Calibrate before use: Improving few-shot performance of language models"

this paper points out that we might have uncanny biases when use in-context learning

Prompt (test input not shown)	**Acc.**
Review: the whole thing 's fairly lame , making it par for the course for disney sequels . Answer: Negative Review: this quiet , introspective and entertaining independent is worth seeking . Answer: Positive	88.5%
Review: this quiet , introspective and entertaining independent is worth seeking . Answer: Positive Review: the whole thing 's fairly lame , making it par for the course for disney sequels . Answer: Negative	51.3%

accuracy scores fluctuate with order of examples in in-context learning

same in-context examples but different order,
but the acc. dropped dramatically

Zhao et al. (2021)

Accuracy variance under spurious prompt variation

"Calibrate before use: Improving few-shot performance of language models"

- accuracy of in-context learning performance from: `influential factors`
  - prompt format
  - training examples
  - order of examples
- identified sources of variation
  - majority label bias
    - prefer answer categories that appear frequently in the example list
  - recency label bias
    - prefer answer categories that appear late in the example list
  - common token bias `almost impossible to get rid of`
    - prefer answer tokens / strings that are common in the training data

`this is not the previous `calibration` which is a question about model itself after pre-training, here is sth we do in order to get rid of the system biases`

Calibration

- $\mathbf{p}_\varnothing$: probabilities of answer categories under a null prompt

Input:	Subpar acting.	Sentiment:	Negative
Input:	Beautiful film.	Sentiment:	Positive
Input:	N/A	Sentiment:	

- define affine transformation

$$F(\mathbf{x}) = \text{soft-max}\left(W\mathbf{x} + \mathbf{b}\right)$$

`train a linear reg model to calculate the bias measured`

- set diagonal matrix W & vector $\mathbf{b}$ so that $F\left(\mathbf{p}_\varnothing\right)$ is max. entropy
  - infinitely many solutions !!!
- choose option:

$$\arg\max_i F\left(\mathbf{x}_i\right)$$

Zhao et al. (2021)

Calibration improves accuracy

could improve.

Dataset	LM	0-shot		1-shot		4-shot		8-shot	
		Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
Text Classification									
AGNews	2.7B	44.7 $_{0.0}$	**63.2** $_{0.0}$	33.0 $_{5.1}$	**59.6** $_{6.4}$	43.3 $_{8.3}$	**71.1** $_{8.5}$	50.8 $_{7.8}$	**72.7** $_{5.8}$
	175B	43.9 $_{0.0}$	**73.9** $_{0.0}$	62.1 $_{6.3}$	**77.1** $_{3.8}$	61.0 $_{10.9}$	**85.9** $_{1.3}$	79.1 $_{2.6}$	**84.3** $_{2.5}$
TREC	2.7B	31.0 $_{0.0}$	**38.8** $_{0.0}$	24.3 $_{6.4}$	**36.8** $_{7.7}$	25.8 $_{11.5}$	**38.6** $_{13.2}$	29.3 $_{8.0}$	**44.3** $_{11.4}$
	175B	47.4 $_{0.0}$	**57.4** $_{0.0}$	57.7 $_{6.0}$	**75.7** $_{1.4}$	60.2 $_{7.6}$	**69.7** $_{1.4}$	45.6 $_{4.0}$	**66.9** $_{6.5}$
CB	2.7B	44.6 $_{0.0}$	**50.0** $_{0.0}$	**33.8** $_{16.6}$	33.0 $_{7.3}$	43.5 $_{11.9}$	**54.2** $_{4.7}$	43.9 $_{8.4}$	**53.0** $_{7.7}$
	175B	30.4 $_{0.0}$	**48.2** $_{0.0}$	50.9 $_{6.7}$	**51.8** $_{7.2}$	45.2 $_{19.4}$	**60.7** $_{6.7}$	59.6 $_{11.3}$	**65.0** $_{7.9}$
RTE	2.7B	44.8 $_{0.0}$	**49.5** $_{0.0}$	49.6 $_{2.9}$	**50.4** $_{2.7}$	44.0 $_{1.4}$	**54.5** $_{4.7}$	49.2 $_{1.9}$	**54.8** $_{2.8}$
	175B	**57.8** $_{0.0}$	**57.8** $_{0.0}$	**62.9** $_{2.7}$	62.8 $_{2.3}$	58.7 $_{11.9}$	**60.4** $_{8.1}$	**66.2** $_{5.8}$	65.5 $_{2.5}$
SST-2	2.7B	57.2 $_{0.0}$	**71.4** $_{0.0}$	67.3 $_{7.9}$	**79.1** $_{8.3}$	59.1 $_{10.2}$	**79.9** $_{7.8}$	54.0 $_{4.3}$	**82.0** $_{5.5}$
	175B	71.6 $_{0.0}$	**75.8** $_{0.0}$	93.3 $_{2.8}$	**94.7** $_{1.4}$	93.6 $_{3.3}$	**94.3** $_{1.0}$	**95.6** $_{1.0}$	95.3 $_{0.7}$
DBPedia	2.7B	36.0 $_{0.0}$	**38.7** $_{0.0}$	25.9 $_{4.4}$	**61.6** $_{2.9}$	61.0 $_{12.8}$	**66.0** $_{7.5}$	72.6 $_{4.5}$	**74.8** $_{5.0}$
	175B	22.0 $_{0.0}$	**59.7** $_{0.0}$	79.3 $_{3.0}$	**85.3** $_{2.2}$	84.6 $_{5.8}$	**86.9** $_{4.0}$	82.3 $_{7.8}$	**86.9** $_{1.9}$
Fact Retrieval									
LAMA	2.7B	14.0 $_{0.0}$	**22.7** $_{0.0}$	29.7 $_{1.8}$	**31.6** $_{1.3}$	35.8 $_{3.8}$	**37.4** $_{3.4}$	**42.5** $_{1.3}$	**42.5** $_{1.4}$
	175B	23.5 $_{0.0}$	**30.1** $_{0.0}$	48.9 $_{2.3}$	**49.0** $_{1.4}$	**62.0** $_{2.4}$	61.8 $_{2.9}$	**63.8** $_{1.0}$	63.6 $_{1.3}$
Information Extraction									
MIT-G	2.7B	5.0 $_{0.0}$	**5.7** $_{0.0}$	26.7 $_{11.4}$	**37.9** $_{5.7}$	53.1 $_{7.8}$	**54.7** $_{6.0}$	59.0 $_{4.7}$	**59.1** $_{4.8}$
	13B	15.0 $_{0.0}$	**18.7** $_{0.0}$	47.3 $_{3.9}$	**52.0** $_{7.9}$	57.9 $_{4.8}$	**58.9** $_{4.0}$	59.0 $_{4.7}$	**59.1** $_{4.8}$
MIT-D	2.7B	46.3 $_{0.0}$	**47.0** $_{0.0}$	42.0 $_{13.0}$	**53.5** $_{13.5}$	73.5 $_{4.9}$	**74.1** $_{5.0}$	**75.3** $_{1.0}$	75.1 $_{1.3}$
	13B	36.3 $_{0.0}$	**38.7** $_{0.0}$	58.6 $_{21.4}$	**72.8** $_{4.0}$	75.4 $_{1.9}$	**75.9** $_{2.1}$	**77.8** $_{0.5}$	**77.8** $_{0.5}$
ATIS-A	2.7B	10.8 $_{0.0}$	**14.0** $_{0.0}$	29.8 $_{12.8}$	**33.1** $_{9.4}$	43.0 $_{26.2}$	**47.3** $_{21.3}$	55.6 $_{5.0}$	**58.8** $_{4.0}$
	13B	49.5 $_{0.0}$	**52.7** $_{0.0}$	69.6 $_{17.4}$	**71.8** $_{17.1}$	67.5 $_{10.4}$	**69.6** $_{13.4}$	63.4 $_{4.6}$	**64.5** $_{4.0}$
ATIS-D	2.7B	6.4 $_{0.0}$	**12.9** $_{0.0}$	42.3 $_{28.8}$	**65.6** $_{20.8}$	75.0 $_{6.7}$	**83.4** $_{4.2}$	81.0 $_{8.8}$	**88.3** $_{3.7}$
	13B	4.0 $_{0.0}$	**5.0** $_{0.0}$	**97.9** $_{0.6}$	95.5 $_{4.6}$	**98.0** $_{0.6}$	97.8 $_{0.7}$	**98.8** $_{0.3}$	**98.8** $_{0.3}$

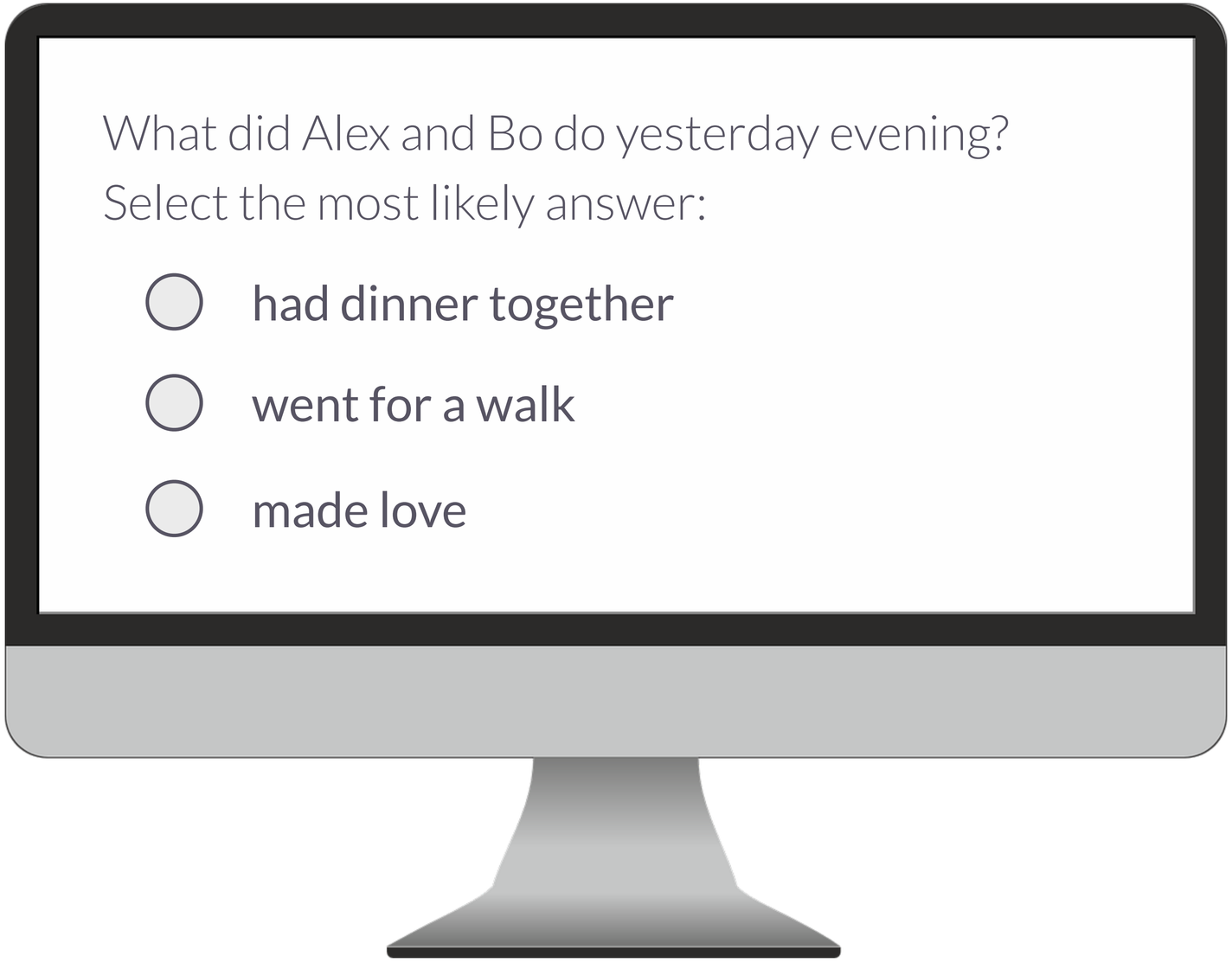
Table 1. **Contextual calibration improves accuracy across a range of tasks.** We show the mean and standard deviation across different choices of the training examples (the prompt format is fixed). The LM column indicates the GPT-3 size (see Appendix A for GPT-2 results). The Baseline column shows the standard approach of greedy decoding (Brown et al., 2020) and *Ours* corresponds to greedy decoding after modifying the output probabilities using contextual calibration. We bold the better result of the baseline and ours. MIT-G, MIT-D, ATIS-A, and ATIS-D indicate the MIT Genre, MIT Director, ATIS Airline, and ATIS Departure Date datasets.

Zhao et al. (2021)

Surface form competition

a problem for multiple-choice task predictions

`one more bias, but there are tons in reality`



Holtzmann et al. (2021)

Domain-conditional correction

officially: “domain-conditional point-wise mutual information” *

```
a scoring rule to get rid of Surface Form Competition
```

choose string option $\mathbf{y}_i$ given task question $\mathbf{x}$ via:

$$\arg\max_i \frac{P(\mathbf{y}_i \mid \mathbf{x})}{P(\mathbf{y}_i \mid \mathbf{x}_{\text{domain}})}$$

where $\mathbf{x}_{\text{domain}}$ is a prompt to indicate the domain

- e.g., just the last word(s) of $\mathbf{x}$

* ironically, this only shares **surface form** with PMI but is not related to the actual concept 🥳

Holtzmann et al. (2021)

Domain-conditional correction

improves accuracy **on top of** calibration

Scoring Functions

Probability (LM)

$$\underset{i}{\text{argmax}}\ P(\mathbf{y}_i|\mathbf{x})$$

Average Log-Likelihood (Avg)

$$\arg\max_i \frac{\sum_{j=1}^{\ell_i} P(y_i^j|\mathbf{x},\mathbf{y}^{1\cdots j-1})}{\ell_i}$$

Contextual Calibration (CC)

$$\arg\max_i \mathbf{w}P(\mathbf{y}_i|\mathbf{x}) + \mathbf{b}$$

Domain Conditional PMI (PMI_{DC})

$$\arg\max_i \frac{P(\mathbf{y}_i|\mathbf{x})}{P(\mathbf{y}_i|\mathbf{x}_{\text{domain}})}$$

Multiple Choice Accuracy on GPT-3

Params.	2.7B					6.7B				13B				175B				
	Unc	LM	Avg	PMI_{DC}	CC	Unc	LM	Avg	PMI_{DC}	Unc	LM	Avg	PMI_{DC}	Unc	LM	Avg	PMI_{DC}	CC
COPA	54.8	68.4	68.4	**74.4**	-	56.4	75.8	73.6	**77.0**	56.6	79.2	77.8	**84.2**	56.0	85.2	82.8	**89.2**	-
SC	50.9	66.0	68.3	**73.1**	-	51.4	70.2	73.3	**76.8**	52.0	74.1	77.8	**79.9**	51.9	79.3	83.1	**84.0**	-
HS	31.1	34.5	**41.4**	34.2	-	34.7	40.8	**53.5**	40.0	38.8	48.8	**66.2**	45.8	43.5	57.6	**77.2**	53.5	-
R-M	22.4	37.8	42.4	**42.6**	-	21.2	43.3	45.9	**48.5**	22.9	49.6	50.6	**51.3**	22.5	55.7	**56.4**	55.7	-
R-H	21.4	30.3	32.7	**36.0**	-	22.0	34.8	36.8	**39.8**	22.9	38.2	39.2	**42.1**	22.2	42.4	43.3	**43.7**	-
ARC-E	31.6	**50.4**	44.7	44.7	-	33.5	**58.2**	52.3	51.5	33.8	**66.2**	59.7	57.7	36.2	**73.5**	67.0	63.3	-
ARC-C	21.1	21.6	25.5	**30.5**	-	21.8	26.8	29.8	**33.0**	22.3	32.1	34.3	**38.5**	22.6	40.2	43.2	**45.5**	-
OBQA	10.0	17.2	27.2	**42.8**	-	11.4	22.4	35.4	**48.0**	10.4	28.2	41.2	**50.4**	10.6	33.2	43.8	**58.0**	-
CQA	15.9	33.2	36.0	**44.7**	-	17.4	40.0	42.9	**50.3**	16.4	48.8	47.9	**58.5**	16.3	61.0	57.4	**66.7**	-
BQ	**62.2**	58.5	58.5	53.5	-	37.8	**61.0**	**61.0**	**61.0**	**62.2**	61.1	61.1	60.3	37.8	62.5	62.5	**64.0**	-
RTE	47.3	48.7	48.7	**51.6**	49.5	52.7	**55.2**	**55.2**	48.7	52.7	52.7	52.7	**54.9**	47.3	56.0	56.0	**64.3**	57.8
CB	08.9	51.8	51.8	**57.1**	50.0	08.9	33.9	33.9	**39.3**	08.9	**51.8**	**51.8**	50.0	08.9	48.2	48.2	**50.0**	48.2
SST-2	49.9	53.7	53.76	**72.3**	71.4	49.9	54.5	54.5	**80.0**	49.9	69.0	69.0	**81.0**	49.9	63.6	63.6	71.4	**75.8**
SST-5	18.1	20.0	20.4	**23.5**	-	18.1	27.8	22.7	**32.0**	18.1	18.6	**29.6**	**19.1**	17.6	27.0	27.3	**29.6**	-
AGN	25.0	69.0	69.0	**67.9**	63.2	25.0	**64.2**	**64.2**	57.4	25.0	69.8	69.8	**70.3**	25.0	**75.4**	**75.4**	74.7	73.9
TREC	13.0	29.4	19.2	**57.2**	38.8	22.6	30.2	22.8	**61.6**	22.6	**34.0**	21.4	32.4	22.6	47.2	25.4	**58.4**	57.4

Table 1: Comparison of scoring algorithms when using GPT-3 for zero-shot inference on multiple choice questions.

Holtzmann et al. (2021)

Taking stock: bias corrections

for classification tasks

- biases in predicted category probabilities from
  - unequal length
  - properties of the prompt (e.g., in k-shot learning)
  - unequal numbers of (near) synonyms
    - surface form competition

Machine Psychology

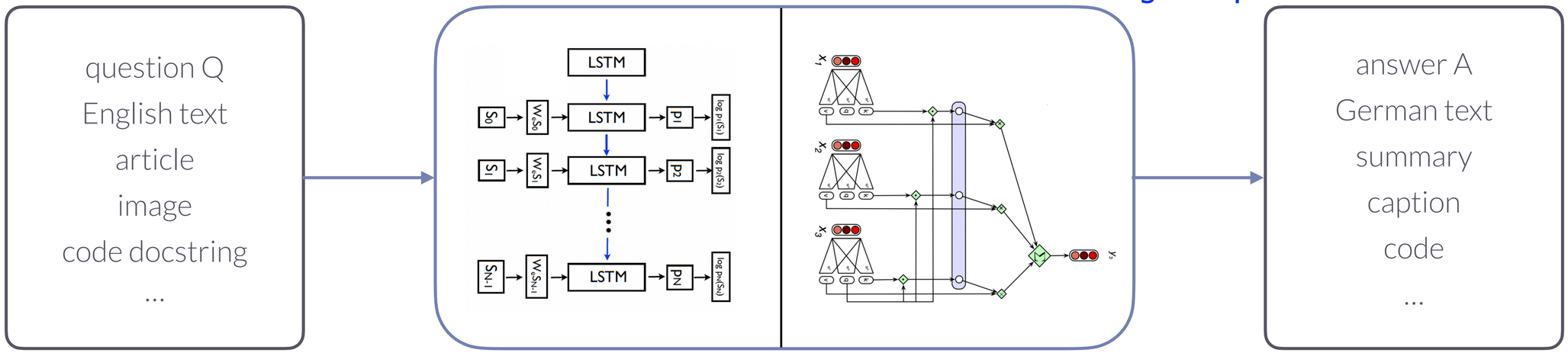
Benchmarks vs. machine psychology

benchmark is interested in whether the systems get the gold standard, but in MP gold standard might not exist

- **benchmarks** (recap)
  - systematic assessment of (emergent) capability
  - streamlined comparison of different systems
    - e.g., **scaling** (performance boost from size *ceteris paribus*)
  - track advancement in the field
  - identify strengths and weaknesses
  - feedback for future work
  - define “what the community cares about”

- **machine psychology**
  - systematic assessment of (emergent) capability
  - studying ML-systems as a new natural kind
    - which kinds of (human-relevant) capabilities are present?
  - using experimental methods inspired from behavioral psychology

irrespective of application purposes,
more natural science,
focus on single questions



input **output**

Machine psychology

"Using cognitive psychology to understand GPT-3"

A Cognitive Psychology View on GPT-3. The core idea behind our approach is to treat GPT-3 as a participant in a psychological experiment. We believe that using such experiments to probe the abilities of large language models has considerable advantages compared to already existing evaluation protocols. In particular, these experiments have been carefully designed to detect various cognitive biases or to disentangle different ways of how a task can be solved. They, therefore, allow us to go beyond the mere performance-based analyses that have been the focus of prior work (10). This is important for two reasons. First, the latest generation of language models is already able to perform above the human level in the majority of tasks from standard benchmark datasets (11, 12), making purely performance-based evaluation less meaningful as time progresses.

benchmark design

Binz & Schulz (2023)

Causal reasoning

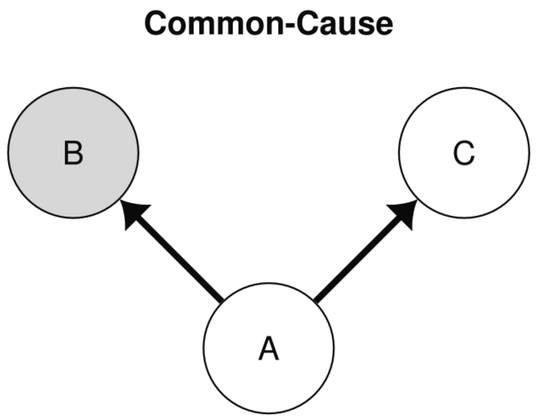
use the same task for human for machines.
A translation from human experiment to machine experiment.

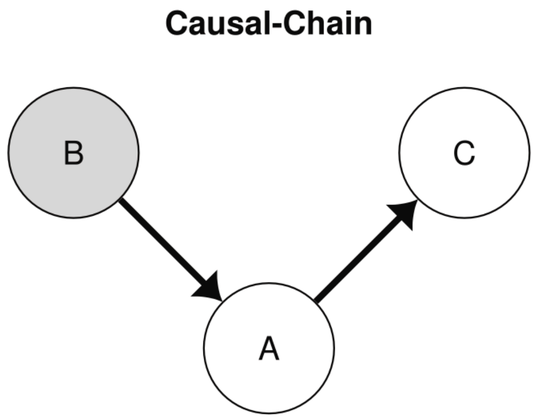
You have previously observed the following chemical substances in different wine casks:
- Cask 1: substance A was present, substance B was present, substance C was present.
- Cask 2: substance A was present, substance B was present, substance C was present.
[...]
- Cask 20: substance A was absent, substance B was absent, substance C was absent.

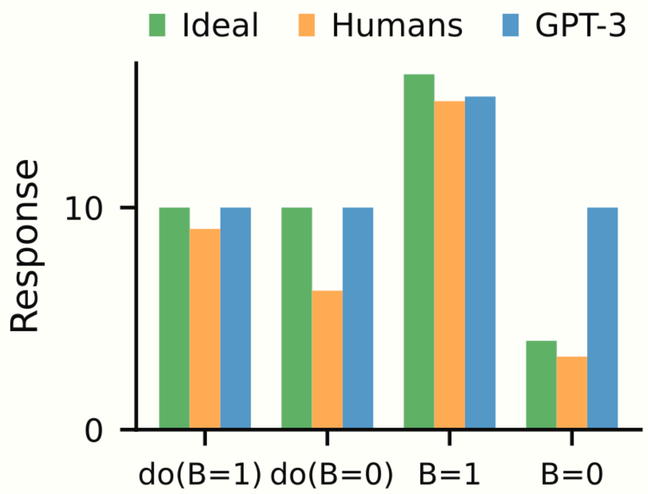
You have the following additional information from previous research:
- Substance A likely causes the production of substance B.
- Substance A likely causes the production of substance C.

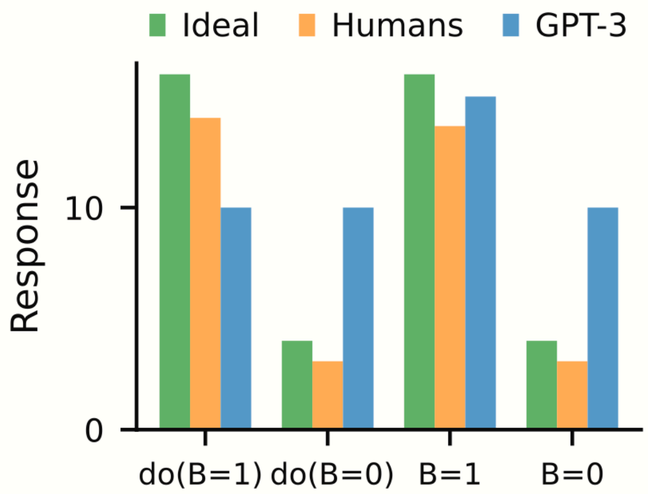
Imagine that you test 20 new casks in which you have manually added substance B.

Q: How many of these new casks will contain substance C on average?
A: [insert] casks.









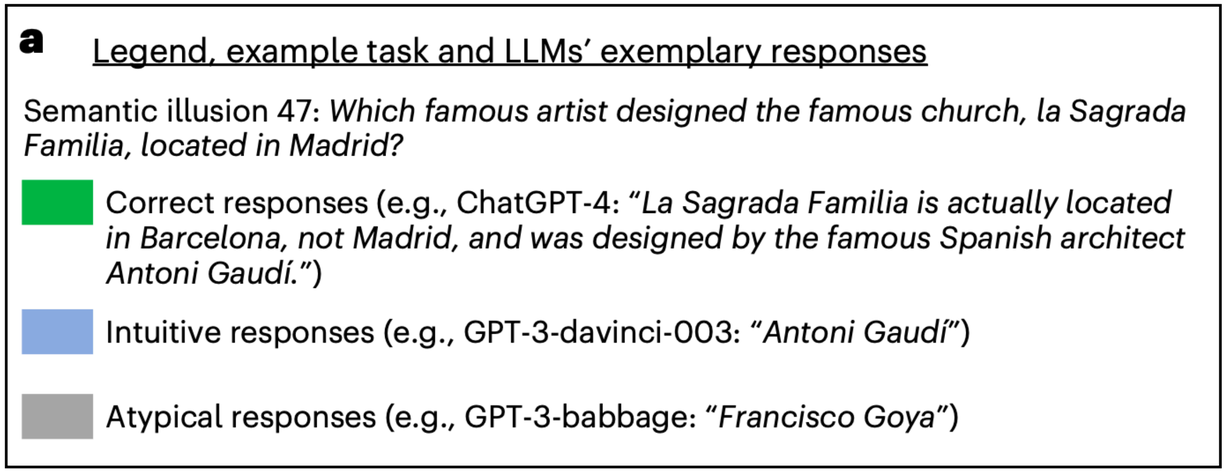
Binz & Schulz (2023)

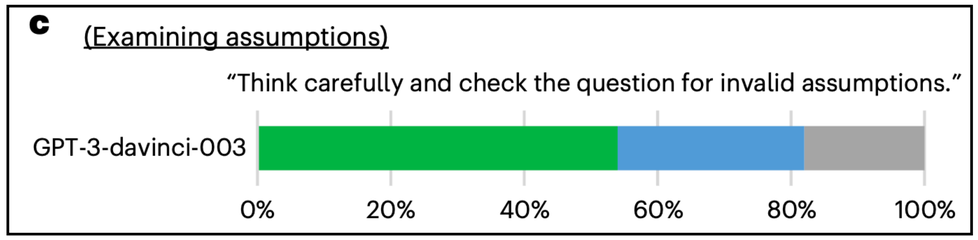
Semantic illusions

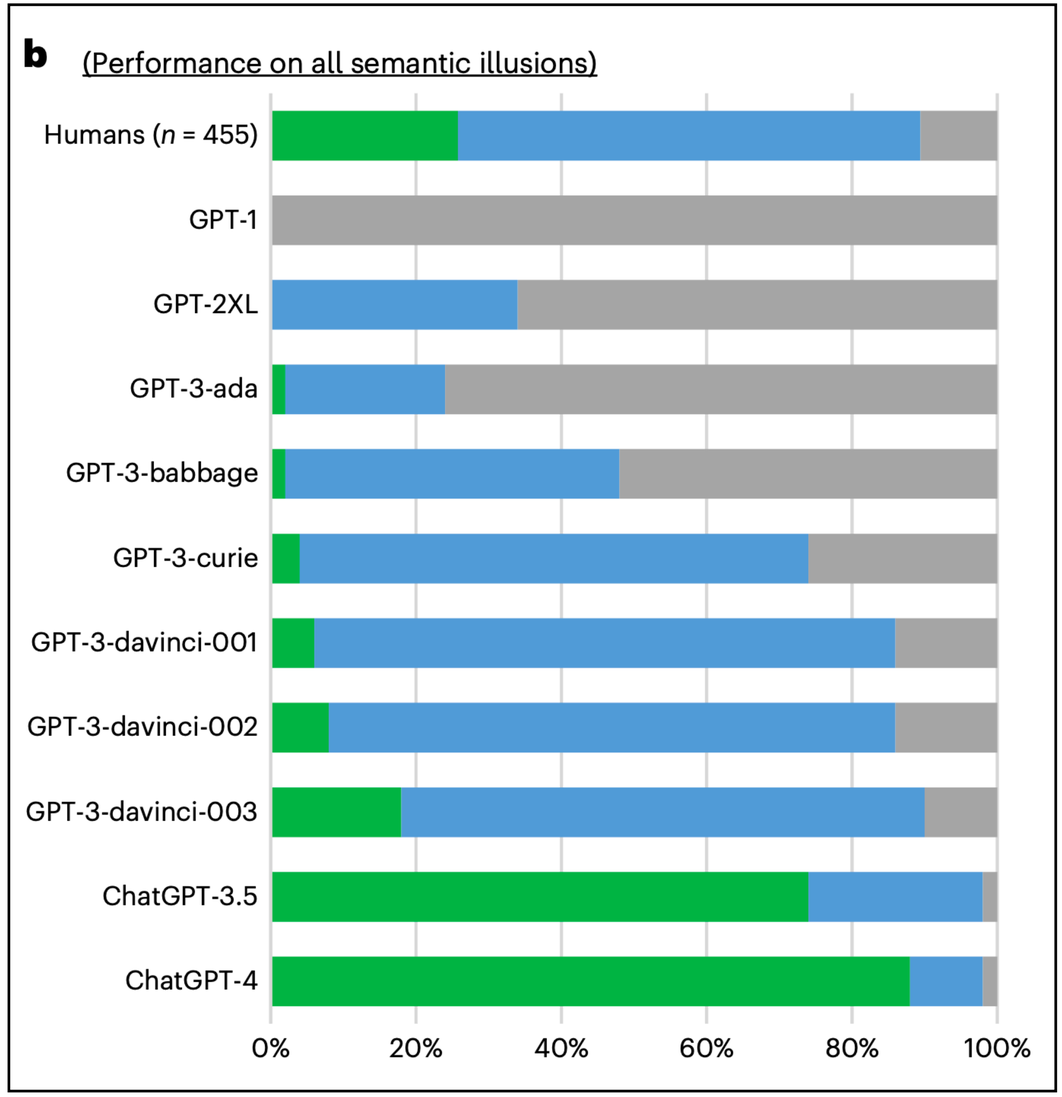
语义幻觉：含有伪装错误的问题，旨在触发直觉但错误的系统响应

- semantic illusions (Erickson & Mattson 1981)
  - questions containing a disguised error aimed at triggering an intuitive but incorrect system 1 response
  - examples:
    - How many animals of each kind did Moses take on the Ark?
    - Who is the dictator of South Korea?

it's Noah not Moses
It's DPRK not SK







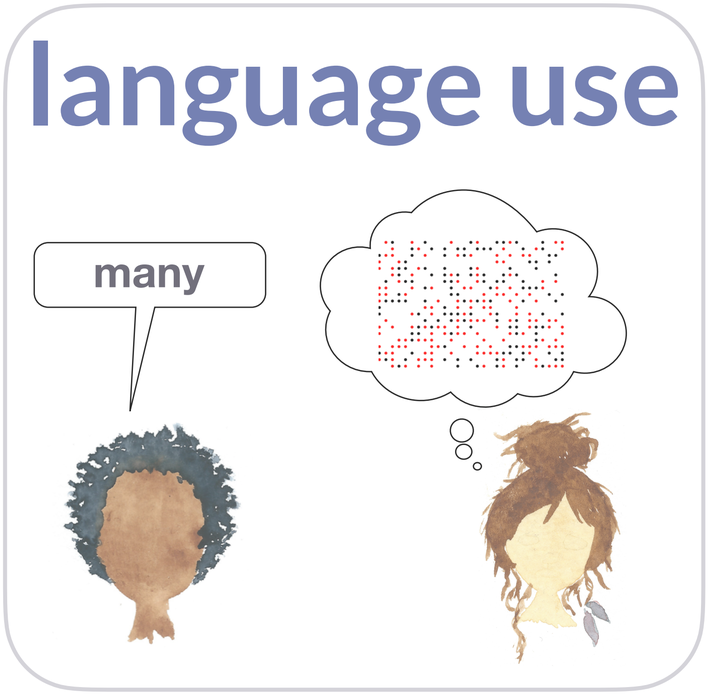
Hagendorff, Fabi & Kosinski (2023, Nature Computational Science)

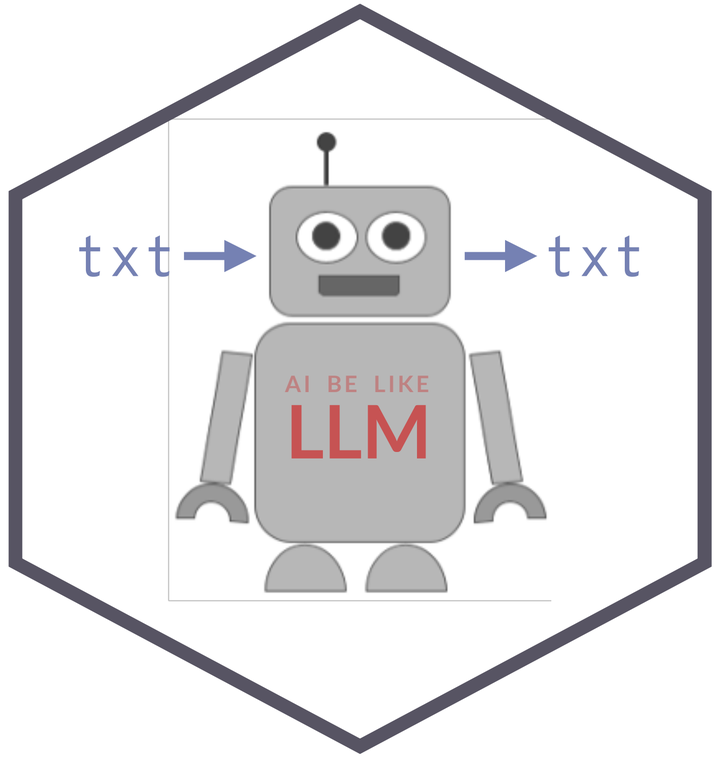
Think break

1. How convincing are the results we saw regarding whether Chat-GPT-X can perform causal reasoning or is susceptible to cognitive biases?
2. What aspects of the methodology used here do you like? What do you see more critically.

Targeted assessment

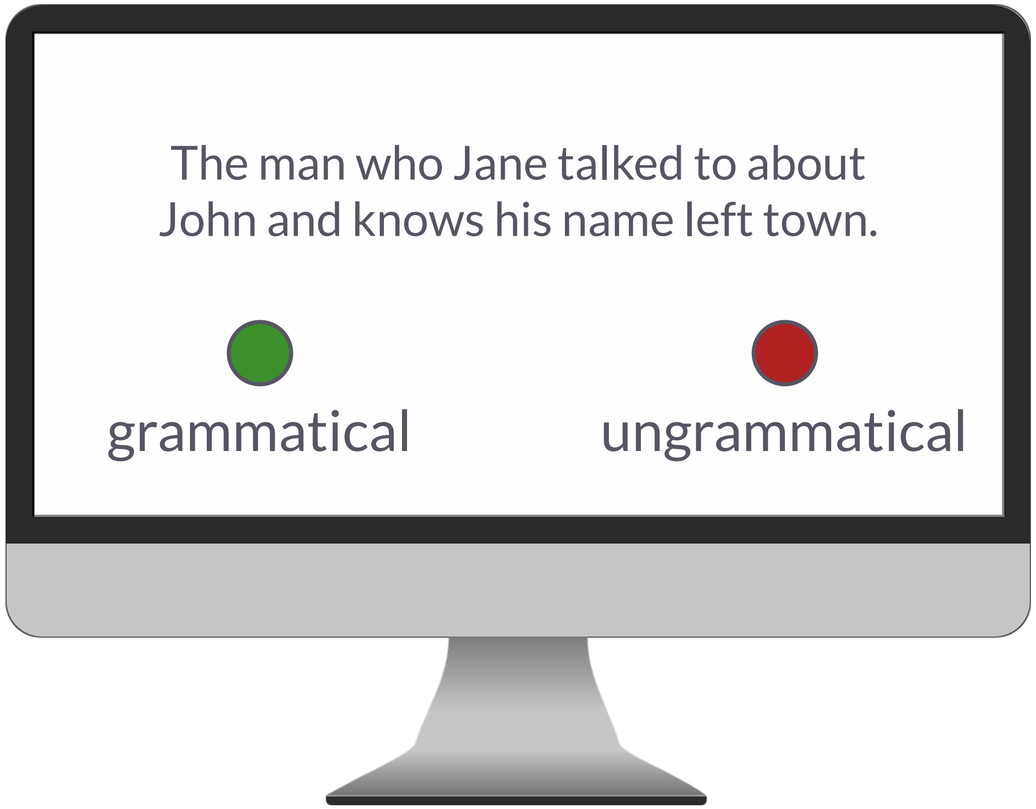
Language: solved!

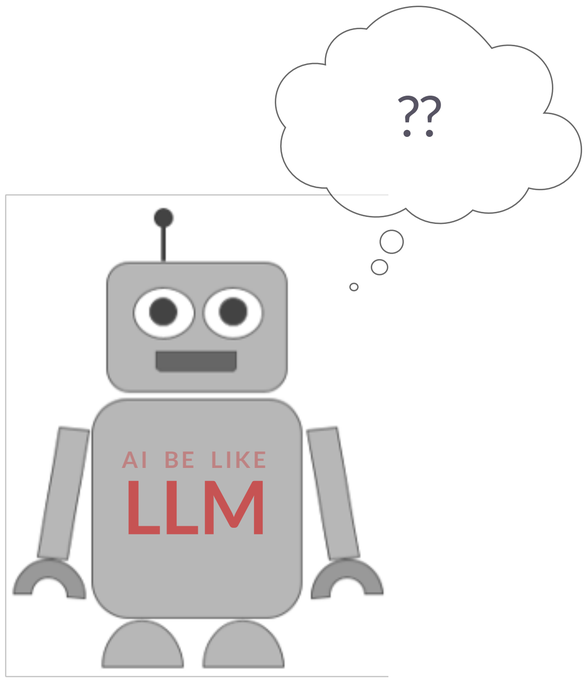




Behavioral experiments

w/ minds & machine





Targeted behavioral assessment

- **research question:**
  - does model M accurately predict
    - human (offline) grammaticality judgements and/ or
    - human (online) processing data?
- **method:**
  - curated test suites (informed by theoretical linguistics & psycholinguistics)
    - e.g., benchmark data set BLiMP (Warstadt et al. 2020)
  - derive model predictions from pre-trained models
  - compare against armchair judgements or actual human data

“Targeted Syntactic Evaluation of LMs”

- three LMs are compared against each other and human data
  - n-gram baseline
  - RNN trained on unannotated data
  - same RNN but with additional CCG supertagging
- test set: ~350k automatically generated sentence pairs
  - generated with a non-recursive context-free grammar
- focus on three phenomena:
  - (i) subject-verb agreement, (ii) reflexive anaphora and (iii) negative polarity
- main findings:
  - performance on training data tracks performance of predicting human grammaticality judgements
  - n-gram baseline < simple RNN < multi-trained RNN

resources

- paper
- code
- video

Marvin & Linzen (2018) EMNLP

Test sentence pairs: SV-Agreement 1

- simple agreement
  - The author laughs.
  - * The author laugh.
  - The authors laugh.
  - * The authors laughs.
- agreement in a sentential complement
  - The bankers knew the officer smiles.
  - * The bankers knew the officer smile.
  - …
- agreement across a prepositional phrase
  - The farmers near the parents smile.
  - * The farmers near the parents smiles.
  - …

Marvin & Linzen (2018) EMNLP

Test sentence pairs: SV-Agreement 2

- agreement across a subject relative clause
  * The officers that love the skater smile.
  * The officers that love the skater smiles.
  …
- short VP coordination
  * The senator smiles and laughs.
  * The senator smiles and laugh.
  …
- long VP coordination
  * The manager writes a letter every day and likes sweets.
  * The manager writes a letter every day and like sweets.
  …

Marvin & Linzen (2018) EMNLP

Test sentence pairs: Agreement in object relative clauses

more difficult: model would need to tell two subjects apart

- agreement across object relative clauses
  - The farmer that the parents love swims.
  - * The farmer that the parents love swim.
  - The farmers that the parent loves swim.
  - * The farmers that the parent loves swims.
- agreement within object relative clauses
  - The farmer that the parents love swims.
  - * The farmer that the parents loves swims.
  - The farmers that the parent loves swim.
  - * The farmers that the parent love swim.

Marvin & Linzen (2018) EMNLP

Test sentence pairs: Agreement in object relative clauses

more difficult: model would need to tell two subjects apart

- simple reflexive
  - The senators embarrassed themselves.
  - * The senators embarrassed herself.
  - …
- reflexive in a sentential complement
  - The bankers thought the pilot embarrassed herself.
  - * The bankers thought the pilot embarrassed themselves. gender neutral?
  - …
- reflexive across an object relative clause
  - The manager that the architects like doubted herself.
  - * The manager that the architects like doubted themselves. gender neutral?
  - …

Marvin & Linzen (2018) EMNLP

Test sentence pairs: Negative polarity

- simple NPI
  - No students have ever lived here.
  - * Most students have ever lived here.
- NPI across a relative clause
  - No authors the guards like have ever been famous.
  - * The authors no guards like have ever been famous.

Marvin & Linzen (2018) EMNLP

Human data

- 100 participants (MTurk)
- each participant saw 76 pairs of sentences
- on each trial, participants had to choose the grammatical sentence from the pair (forced-choice task)
- 16 participants were excluded due to more than one error on the simple agreement trials

Marvin & Linzen (2018) EMNLP

Think break

1. Given a language model, how would we determine whether the model can or cannot match human grammaticality judgements for any pair of sentences without training the model on the task?
2. If human participants make mistakes, what should we expect an LM to do? Be equally good as humans, or be at ceiling where humans fail to meet the grammatical norm?

Defining grammaticality prediction

- 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:

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

Marvin & Linzen (2018) EMNLP

Results

Marvin & Linzen (2018) EMNLP

	RNN	Multitask	*n*-gram	Humans	# sents
SUBJECT-VERB AGREEMENT:					
Simple	0.94	1.00	0.79	0.96	280
In a sentential complement	0.99	0.93	0.79	0.93	3360
Short VP coordination	0.90	0.90	0.51	0.94	1680
Long VP coordination	0.61	0.81	0.50	0.82	800
Across a prepositional phrase	0.57	0.69	0.50	0.85	44800
Across a subject relative clause	0.56	0.74	0.50	0.88	22400
Across an object relative clause	0.50	0.57	0.50	0.85	44800
Across an object relative (no *that*)	0.52	0.52	0.50	0.82	44800
In an object relative clause	0.84	0.89	0.50	0.78	44800
In an object relative (no *that*)	0.71	0.81	0.50	0.79	44800
REFLEXIVE ANAPHORA:					
Simple	0.83	0.86	0.50	0.96	560
In a sentential complement	0.86	0.83	0.50	0.91	6720
Across a relative clause	0.55	0.56	0.50	0.87	44800
NEGATIVE POLARITY ITEMS:					
Simple	0.40	0.48	0.06	0.98	792
Across a relative clause	0.41	0.73	0.60	0.81	31680

Towards systematic assessment of syntactic generalization

- 10 LMs are compared against each other, of which 5 non-pretrained:
  - n-gram baseline, vanilla LSTM, ordered neurons LSTM, RNNG, GTP-2
- 4 different training set sizes (for non-pretrained models)
  - 1, 5, 14 and 42 million tokens
- test set consists of 34 test suits from 6 “syntactic circuits”
  - (i) garden-path effects, (ii) licensing, (iii) agreement, (iv) center embedding
  - (v) long-distance dependencies, (vi) gross syntactic expectation
- introduce **syntactic generalization (SG) score**
- main findings:
  - dissociation between perplexity and SG score
  - model type has more effect on SG than training data size
  - higher SG scores for models with explicit structural training
  - differences in success on different test suits depends on model type

resources

- paper
- code
- video

Hu et al. (2020), ACL

Syntactic generalization (SG) score

- each test suit has a set of predictions
- SG score for test suit X is the proportion of items in X for which the LM matches all predictions associated with X
- example “garden-path sentences”
  - test item example:
    i. The horse raced past the barn fell …
    ii. The horse ridden past the barn fell …
    iii. The horse which was raced past the barn fell …
    iv. The horse which was ridden past the barn fell …
  - associated predictions:

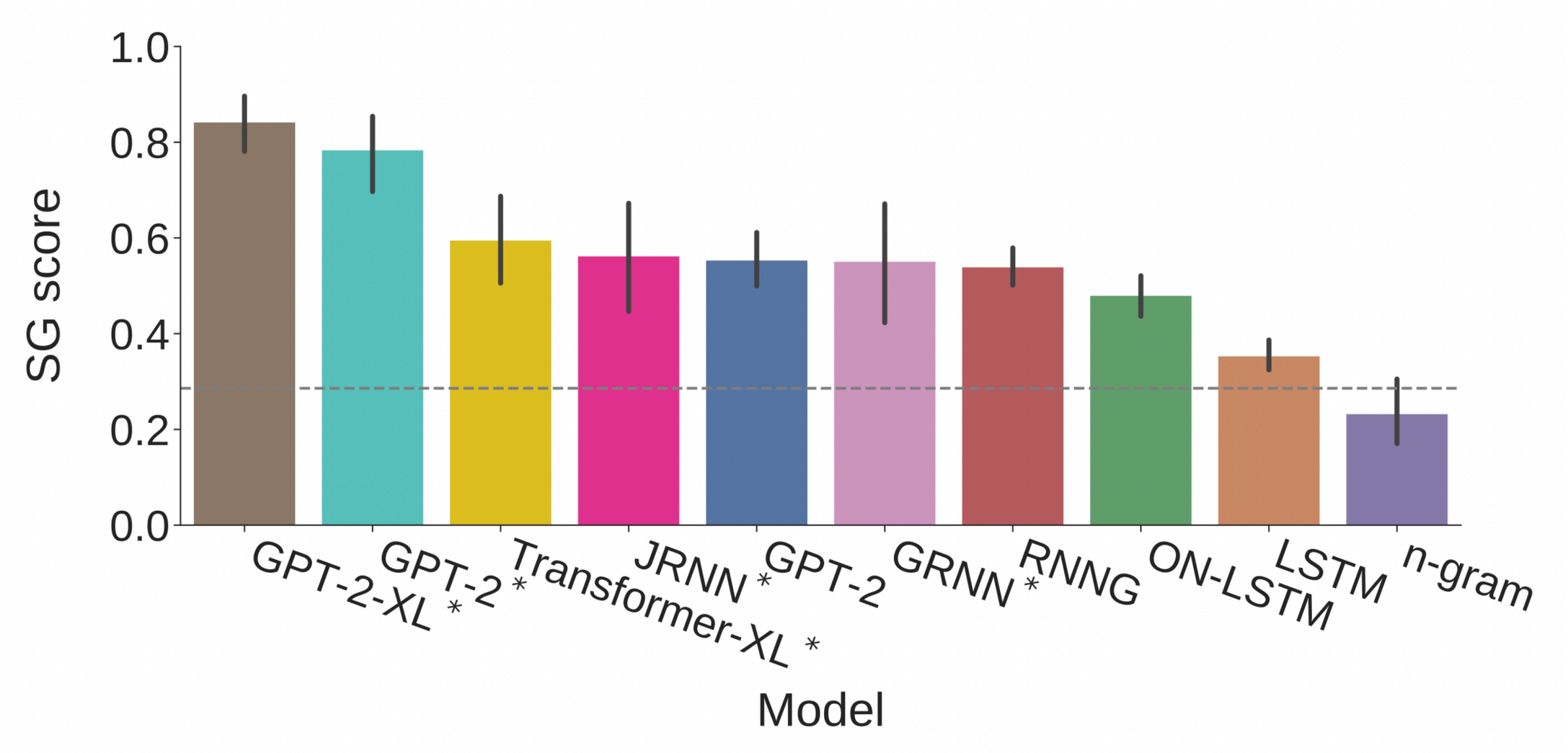
    P(fell | (i)) < P(fell | (ii))

    P(fell | (i)) < P(fell | (iii))

    P(fell | (i)) - P(fell | (ii)) > P(fell | (iii)) - P(fell | (iv))

Hu et al. (2020), ACL

Results: Average SG scores by model type



Hu et al. (2020), ACL

Taking stock: targeted assessment

of LLM's syntactic abilities

- careful design to separate string likelihood from grammaticality
  - e.g., test suits & syntactic generalization scores
- training set size is important, but model architecture (scale) even more

Assessing language processing

Sources of processing difficulty

- limits of working memory

  The dog which the cat which the mouse provoked was chased by barked.

- local ambiguity

  The horse raced past the barn fell.

- interaction w/ semantics & world knowledge

  The cop arrested by the detective was guilty of taking bribes.

Surprisal theory

- **surprisal theory:**
  - $\text{Effort}(w_i, w_{1:i-1}, C) \propto \text{Surprisal}(w_i \mid w_{1:i-1}, C) = -\log P(w_i \mid w_{1:i-1}, C)$
  - compatible with two mechanisms causing processing difficulty:
    - **prediction**: comprehenders actively predict upcoming words; processing difficulty is a form of prediction error
    - **integration**: comprehenders do not actively predict upcoming material, but passive pre-activation leads to easier integration of some material than others
  - empirical evidence for surprisal theory:
    - cloze probability
    - eye-tracked reading
    - self-paced reading
    - EEG during reading
    - maze task

OMG, I didn't see that token coming

Play break

- go try out the **iMaze task** for yourself:
  - follow this link

Targeted Assessment of Incremental Processing in nLMs & Human

- **language models:**
  - JRNN: large-scale RNN using LSTM units & CNN character embeddings
  - GRNN: from Gulordava et al. (2018)
  - GPT-2: version from `lm-zoo` distribution
  - RNNG: average of three RNNGs from Hu et al. (2020)
- **test set:** 16 test suits adapted from Hu et al. (2020)
- **human data** on sentence processing difficulty: decision times from an **iMaze task**

resources

- paper
- code
- video

Wilcox et al. (2021)

Targeted Assessment of Incremental Processing in LMs & Human

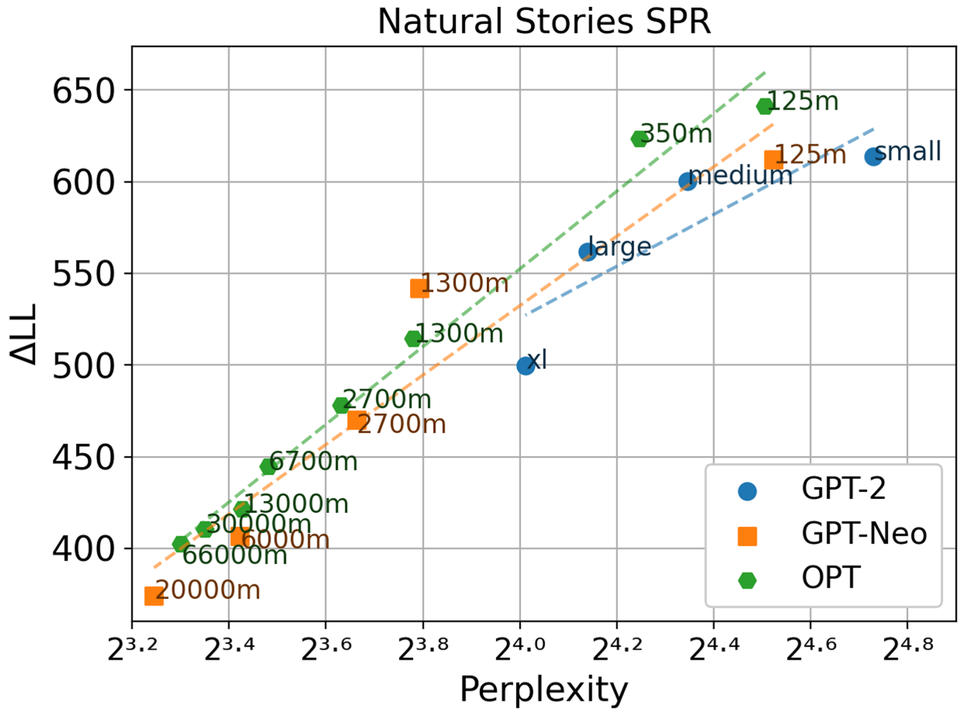
- measure of interest:
  - *qualitative*: accuracy scores (LM prediction vs armchair grammaticality judgements)
  - *quantitative*: degree of slowdown on critical region (LM prediction vs iMaze data)
  - *generalization*: train linear model to map $P_M(w_i \mid w_{1:i-1}) \mapsto RT_{\text{human}}(w_i \mid w_{1:i-1})$ for each w_i not in a critical region, and use it to explain RTs from words in critical regions
- main findings:
  - *qualitative*: LMs predict processing difficulty at regions exactly where humans seem to experience it
  - *quantitative*: LMs are "not surprised enough"
  - *generalization*: LMs routinely underpredict human RTs / surprisal

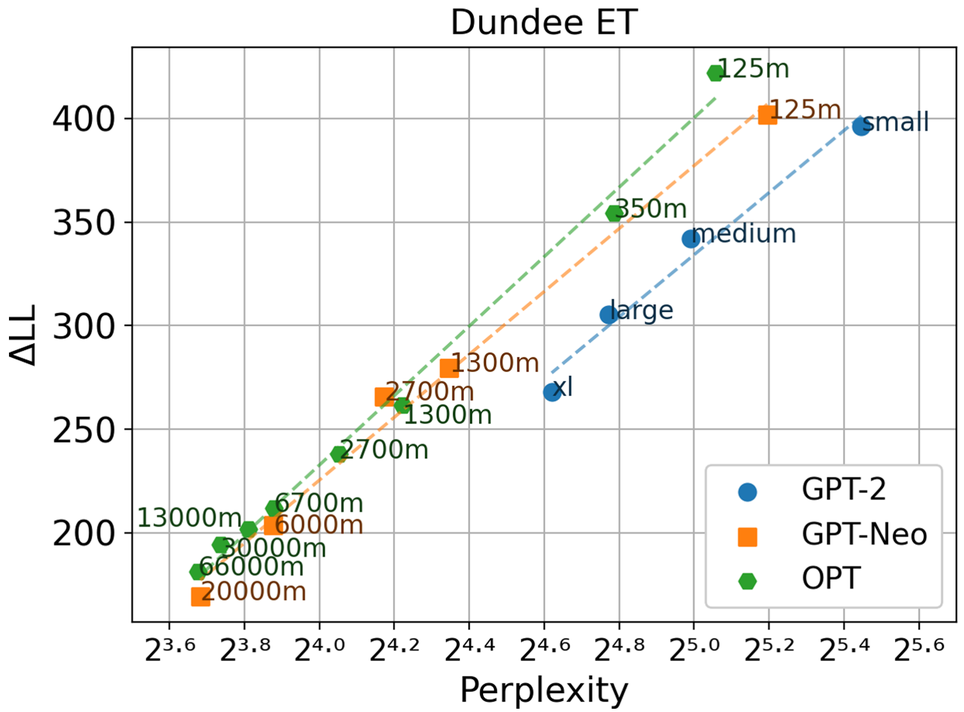
resources

- paper
- code
- video

Wilcox et al. (2021)

The larger and “better” the model, the worse it’s predictions





Oh & Schuler (2023)

Possibly related to miscalibration

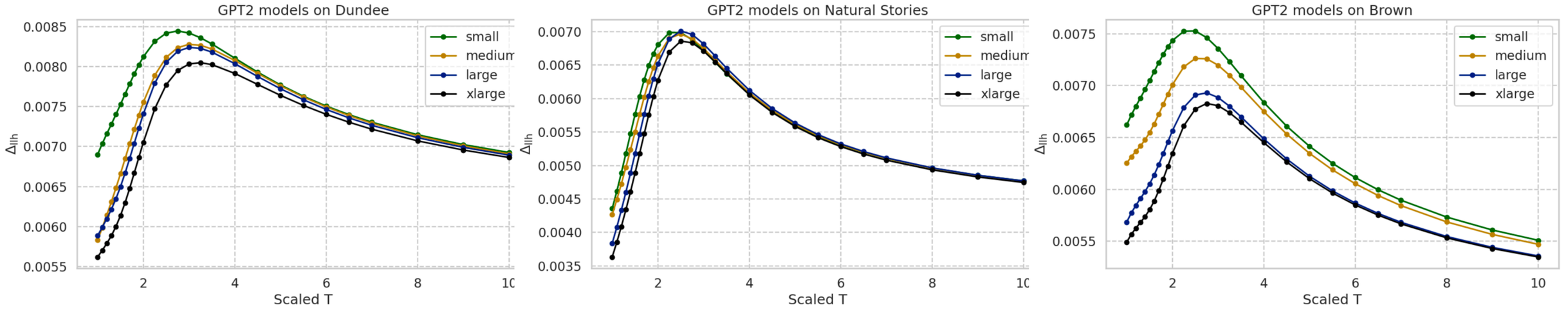


Figure 3: Relationship between Δ_{llh} of GPT-2 models and corresponding temperature. T is scaled from 1.0 to 10.

Liu et al. (2024)

Project idea

- literature survey: measures of human processing difficulty
- investigate the relation between (mis-)calibration and predictive accuracy for human reading times

wanted:

Robust Methods of Machine Psychology

Predictions from an LLM

for multiple choice tasks

Prompt

```
Your task is to play a conversation game. There are three objects
that you and your friend can see. You have to choose a single
word to identify one of the three objects for your friend.

The three objects are:

a blue triangle with spades
a blue circle with stripes
a blue triangle with stripes

Your task is to make your friend pick out the following target
object:

the blue circle with stripes

Which of the following words would you choose:

spades
stripes
circle
triangle

Your answer:

I would choose the word
```

Option scores

spades	26.1
stripes	32.7
circle	18.3
triangle	25.9

argMax choice

spades	0
stripes	**1**
circle	0
triangle	0

softMax choice

spades	0.3
stripes	0.4
circle	0.1
triangle	0.2

Prompt sensitivity

Robustness, consistency

Which of the two following numbers is more common: 17 or 42?

Choose a number: 17

17 = 63.53%
42 = 30.78%
\n = 4.33%
= 0.57%
17 = 0.46%

Total: -0.45 logprob on 1 tokens
(99.67% probability covered in top 5 logits)

Which of the two following numbers is more common: 42 or 17?

Choose a number: 42

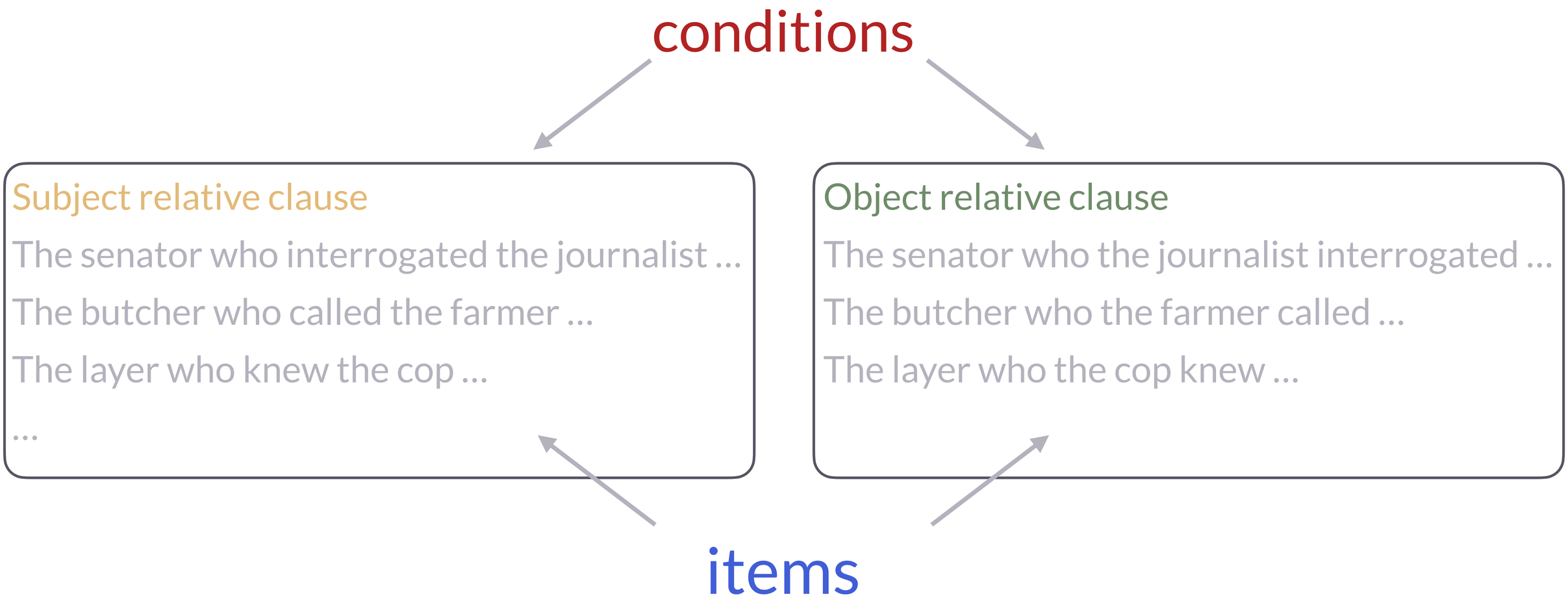
42 = 97.17%
\n = 1.89%
42 = 0.34%
17 = 0.32%
= 0.24%

Total: -0.03 logprob on 1 tokens
(99.96% probability covered in top 5 logits)

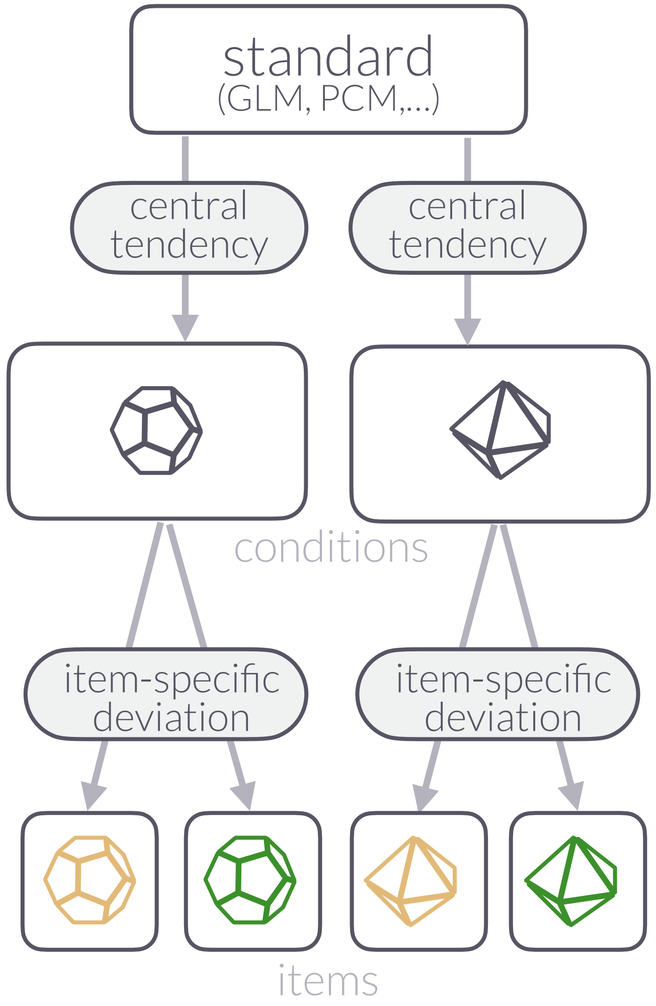
Elazar et al. (2021) “Measuring and Improving Consistency in Pretrained Language Models”

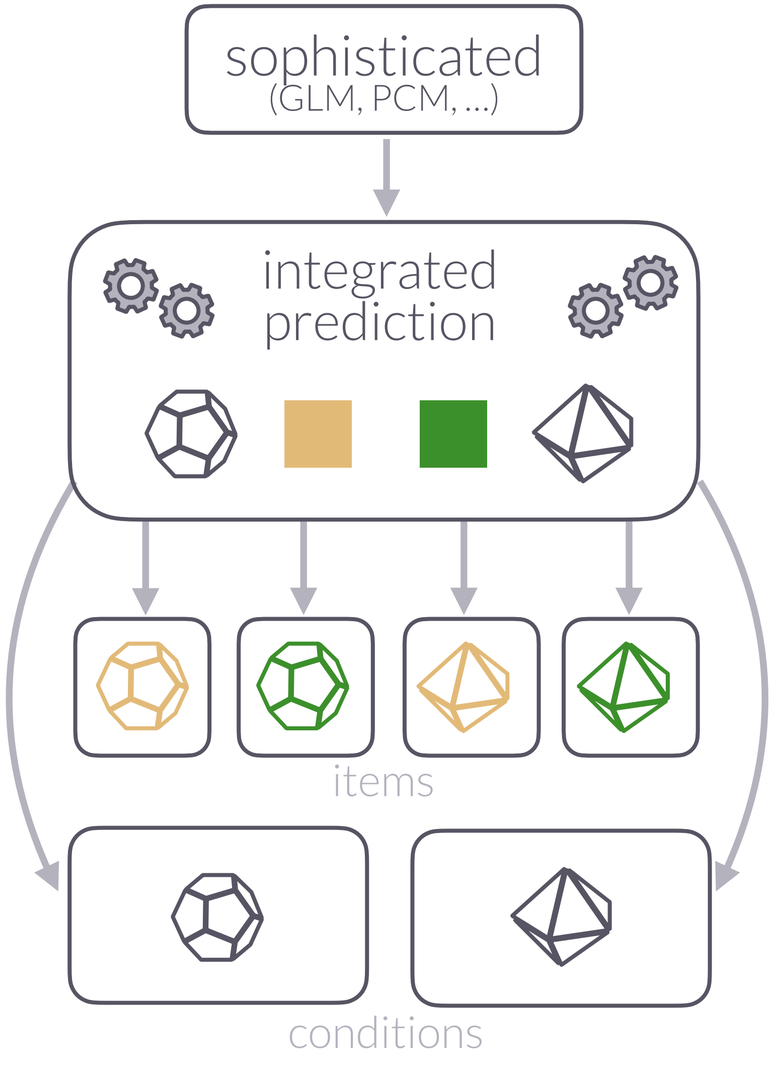
Factorial experimental designs

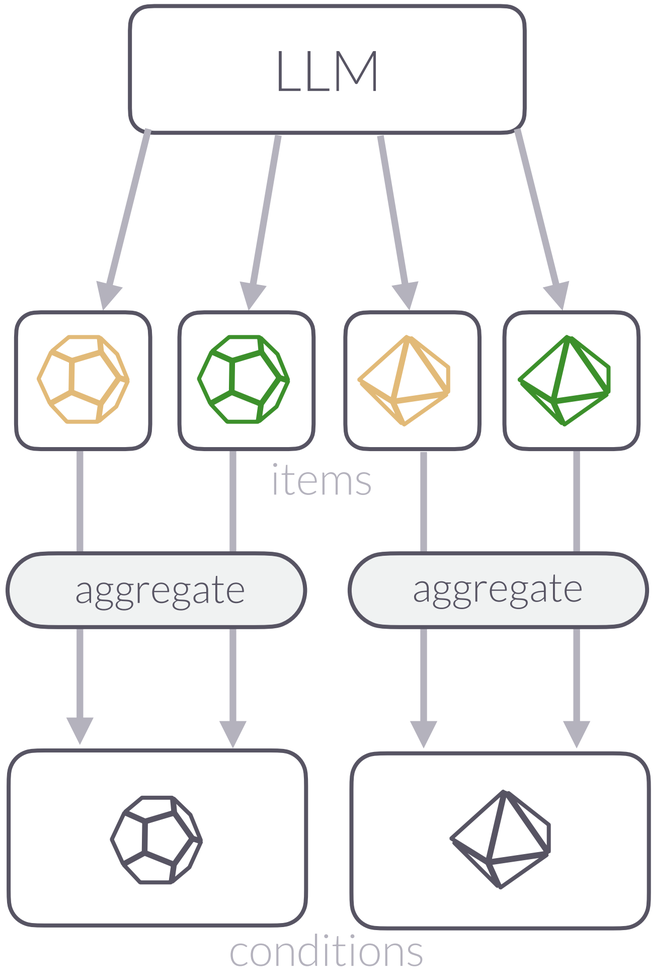
Example: processing relative clauses



Condition- and item-level predictions







Empirical questions

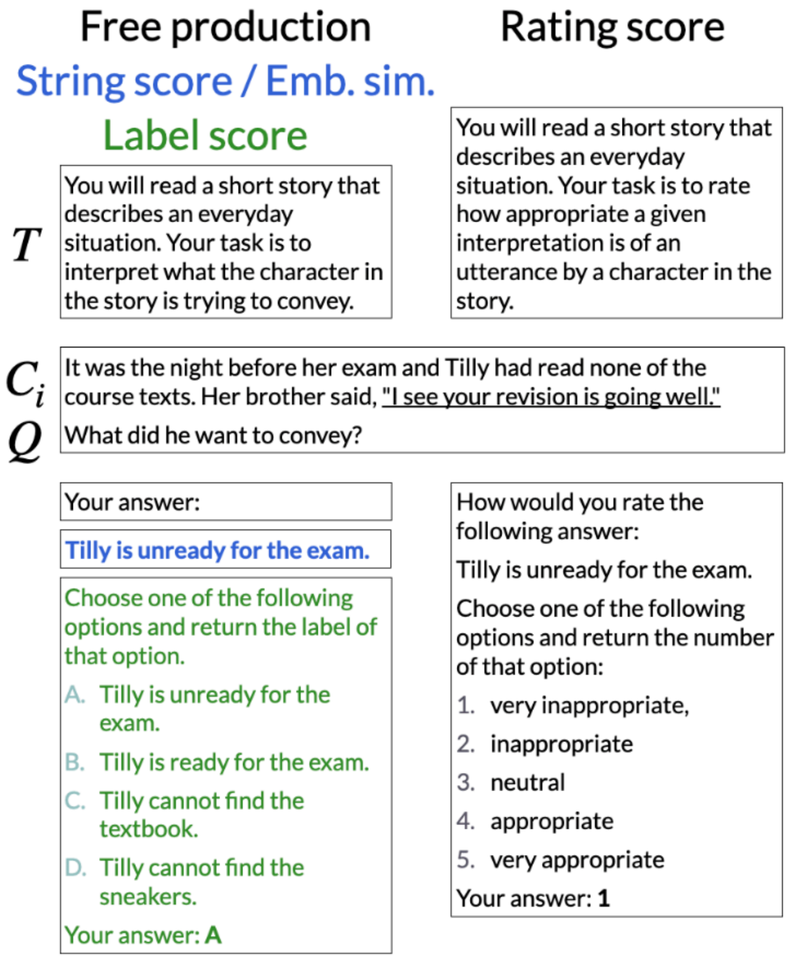
- which scoring scheme gives most accurate and/or human-like predictions?
- variation in LLM's item-level predictions similar to human?
- how to aggregate to obtain condition-level predictions?

Empirical questions

- **which scoring scheme gives most accurate and/or human-like predictions?**
  - **likelihood of human data for full set of choice options**
- variation in LLM's item-level predictions similar to human?
- how to aggregate to obtain condition-level predictions?

Comparing different assessment methods

Tsvilodub, Wang, Grosch & Franke (2024)



(1) option probability $P(o_{ij} \mid I_i)$

(2) avg. opt. prob. $\frac{1}{|o_{ij}|} \sum_l P(w_{ijl} \mid I_i, w_{ij[0\ldots l-1]})$

(3) avg. neg. surprisal $\frac{1}{|o_{ij}|} \sum_l \log P(w_{ijl} \mid I_i, w_{ij[0\ldots l-1]})$

(4) prior corrected opt. prob. $\frac{P(o_{ij} \mid I_i)}{P(o_{ij} \mid I_{i0})}$

(5) surprisal reduction factor $\frac{\log P(o_{ij} \mid I_{i0})}{\log P(o_{ij} \mid I_i)}$

(6) rating aggregation $\sum_{n=1}^{k_i} n \; P(r_n \mid I_{ij})$

(7) embedding similarity $\frac{E(I_i) \cdot E(o_{ij})}{\|E(I_i)\| \cdot \|E(o_{ij})\|}$

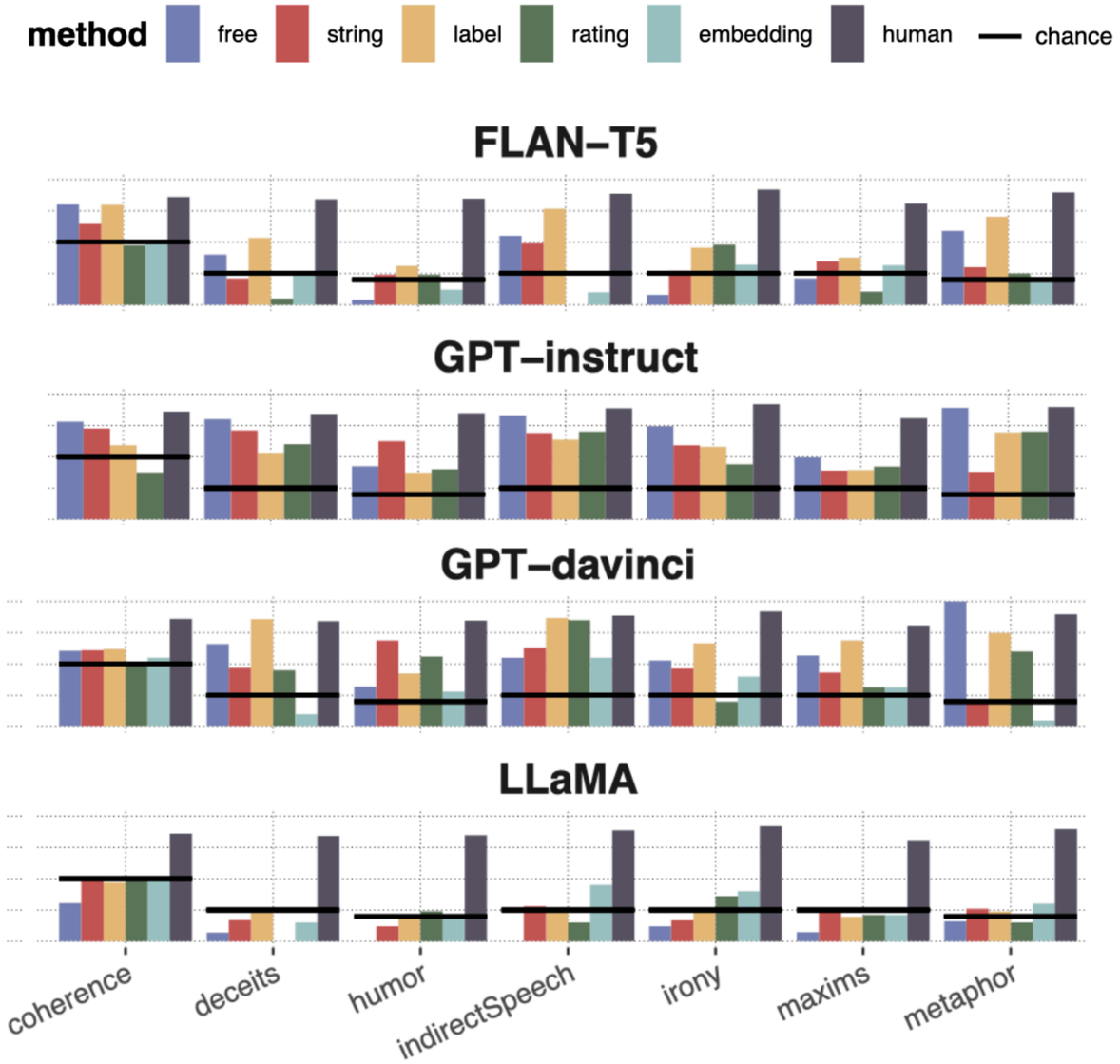
Figure 2: Set of all relevant scores.

building on prior work by Hu & Levy (2023)

Comparing different assessment methods

Tsvilodub, Wang, Grosch & Franke (2024)

- strong variation btw. models & tasks
  - free generation best for GPT-instruct
  - free generation bad for LLaMA2 (7B)
- no clear winner for over all models
- amplifies researcher degrees of freedom



Project idea

- extend this to systematically cover standard benchmarks
- quantify how large the variance is for different LLMs / task types

Empirical questions

- which scoring scheme gives most accurate and/or human-like predictions?
  - likelihood of human data for full set of choice options
- **variation in LLM's item-level predictions similar to human?**
- **how to aggregate to obtain condition-level predictions?**

Human experiment

production condition

Your task is to play a conversation game. There are three objects that you and your friend can see. You have to choose a single word to identify one of the three objects for your friend. The three objects are:

a blue triangle with spades
a blue circle with stripes
a blue triangle with stripes

Your task is to make your friend pick out the following target object:

the blue circle with stripes

Which of the following words do you choose:

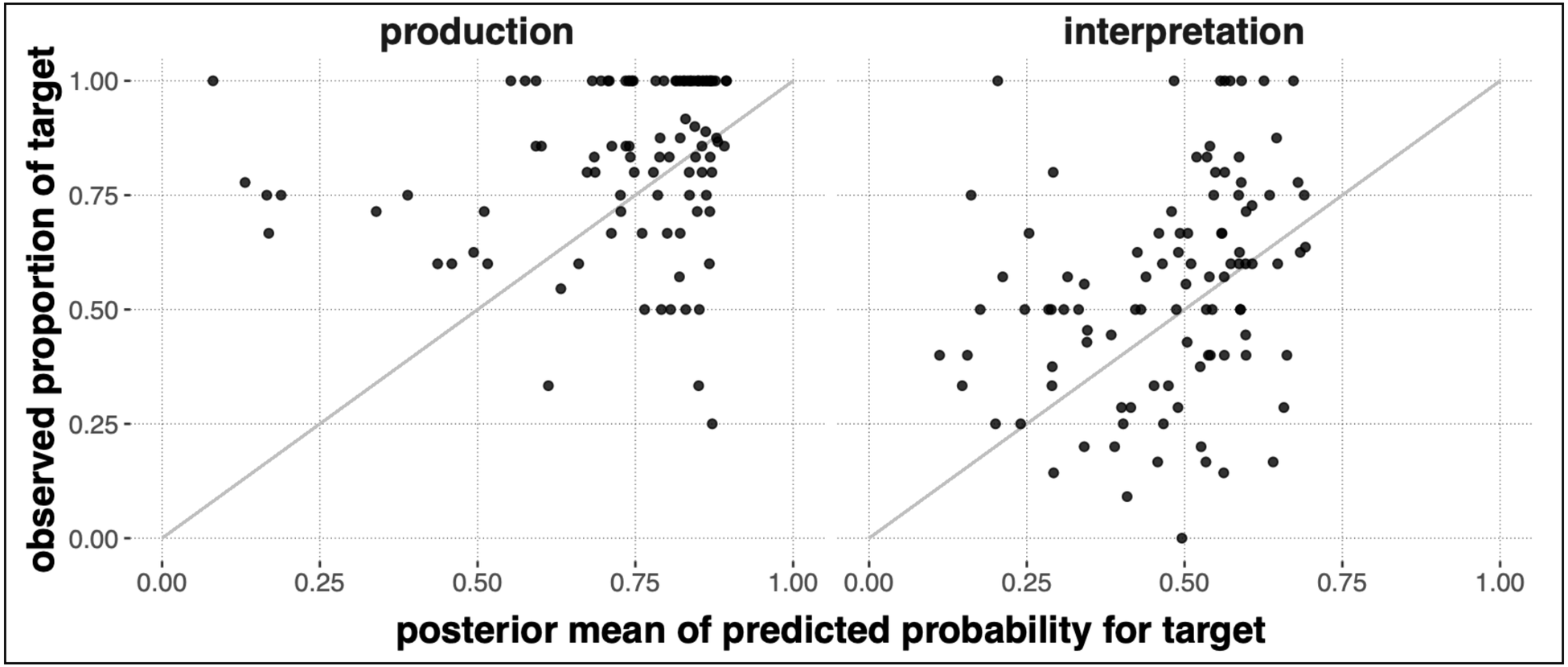
SPADES STRIPES CIRCLE TRIANGLE

LLM prompt

```
Your task is to play a conversation game. There are three objects
that you and your friend can see. You have to choose a single
word to identify one of the three objects for your friend.

The three objects are:

a blue triangle with spades
a blue circle with stripes
a blue triangle with stripes

Your task is to make your friend pick out the following target
object:

the blue circle with stripes

Which of the following words would you choose:

spades
stripes
circle
triangle

Your answer:

I would choose the word
```

Franke, Tsvilodub & Carcassi (in prep)

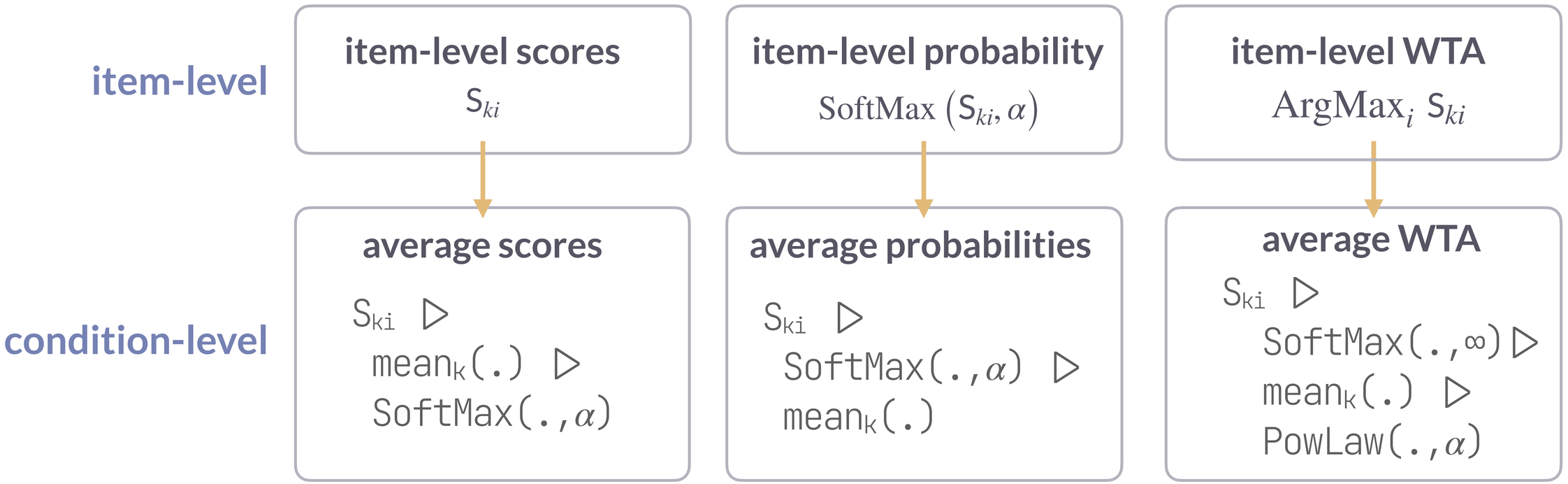
Item-level analysis of LLM predictions



Franke, Tsvilodub & Carcassi (in prep)

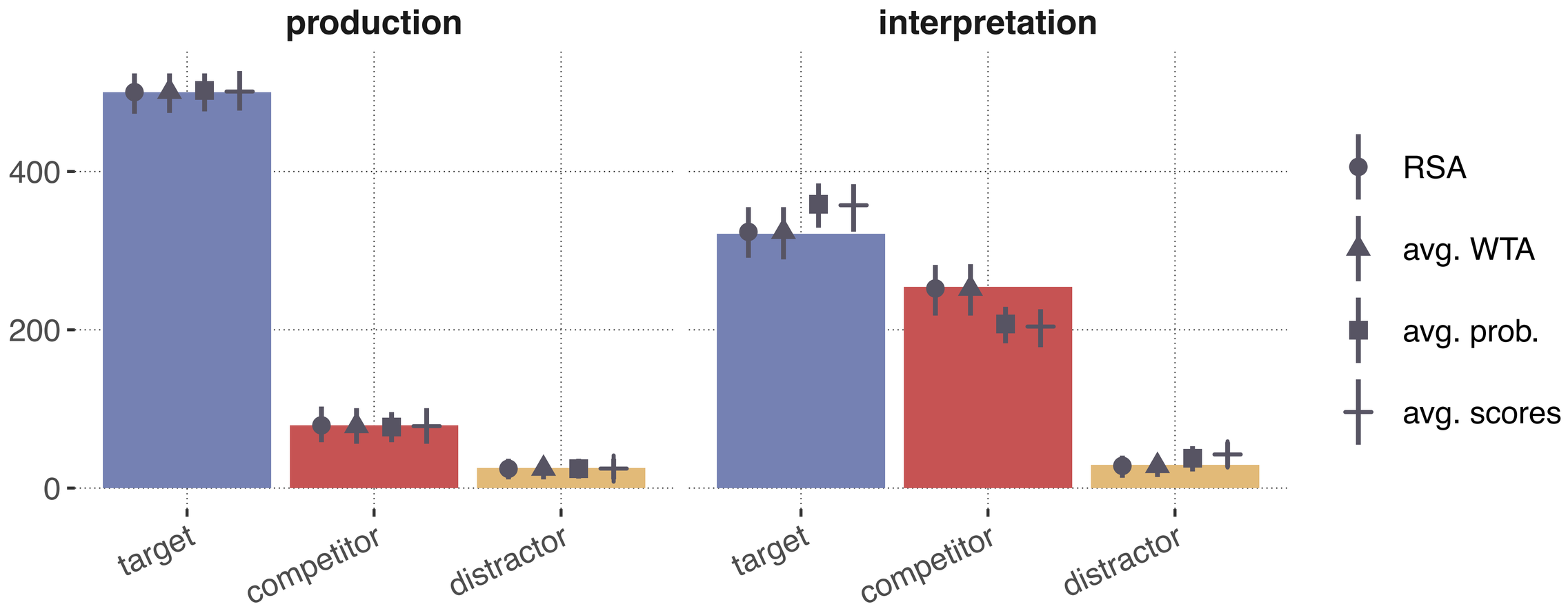
Condition-level predictions

what to aggregate when and how to scale?



Franke, Tsvilodub & Carcassi (2024)

Results: Posterior predictive checks



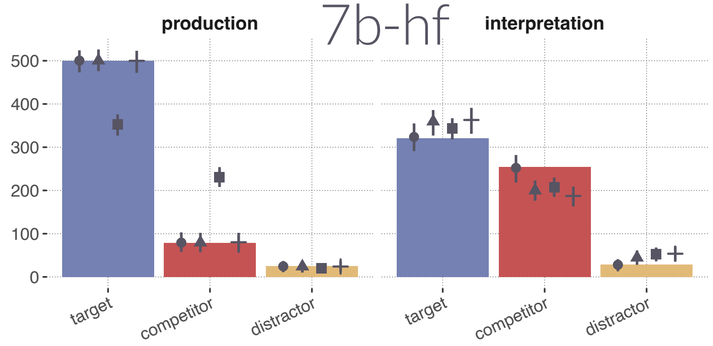
Franke, Tsvilodub & Carcassi (in prep)

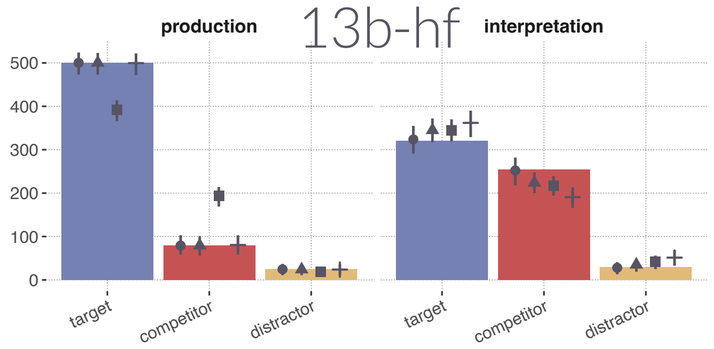
Posterior summaries and predictive checks

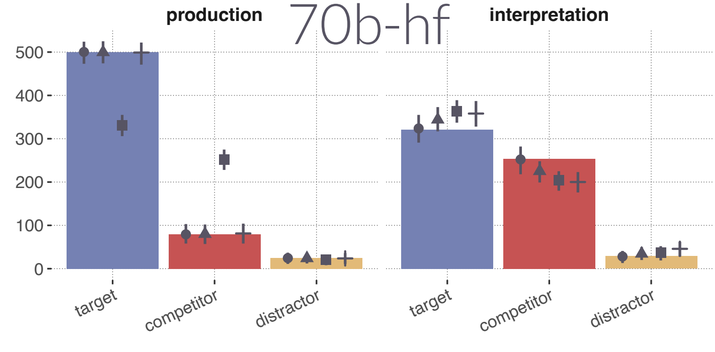
			α			ϵ				
data	model	condition	|95%	mean	95%|	|95%	mean	95%|	Bpppv	
item-level	RSA	production	2.62	3.13	3.69	0.08	0.12	0.16	0.29	
item-level	RSA	interpretation	0.21	0.67	1.08	0.08	0.14	0.19	0.21	
item-level	LLM	production	0.15	0.19	0.24	0.00	0.07	0.16	0.00	*
item-level	LLM	interpretation	0.08	0.11	0.14	0.00	0.09	0.22	0.00	*
cond.-level	RSA	production	2.62	3.14	3.70	0.08	0.12	0.17	0.50	
cond.-level	RSA	interpretation	0.27	0.68	1.13	0.09	0.14	0.19	0.51	
cond.-level	avg. scores	production	0.19	0.22	0.25	0.00	0.03	0.08	0.05	
cond.-level	avg. scores	interpretation	0.15	0.18	0.20	0.00	0.02	0.05	0.00	*
cond.-level	avg. probabilities	production	0.86	1.03	1.22	0.08	0.12	0.17	0.47	
cond.-level	avg. probabilities	interpretation	0.36	0.46	0.59	0.00	0.04	0.10	0.00	*
cond.-level	avg. WTA	production	0.86	1.04	1.22	0.08	0.12	0.17	0.48	
cond.-level	avg. WTA	interpretation	0.15	0.37	0.58	0.03	0.12	0.18	0.49	

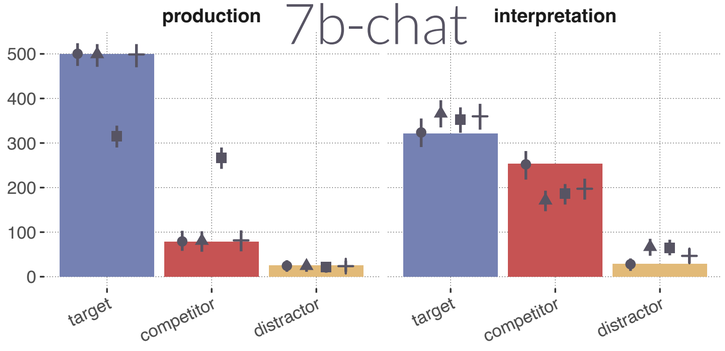
Results: Posterior predictive checks

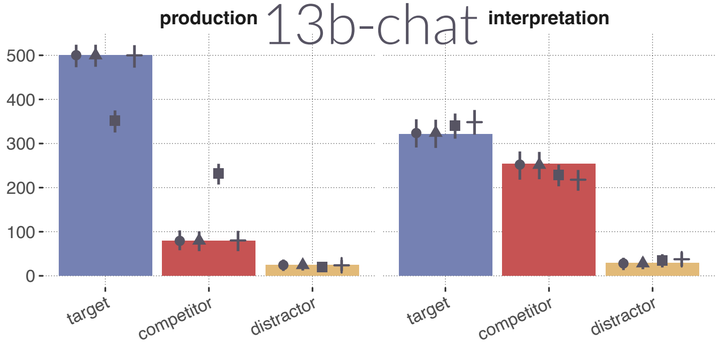
LLaMA variants

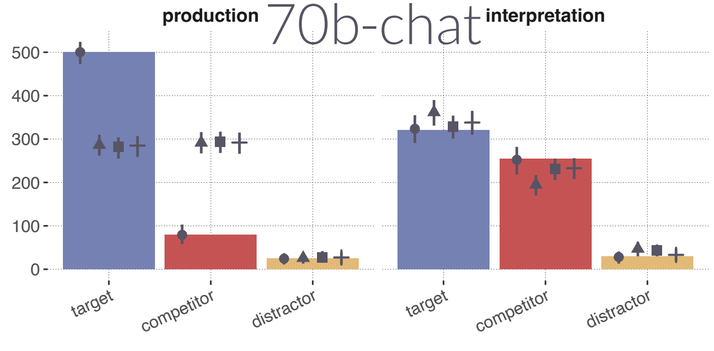












Conclusions

- general:
  LLM's atomic predictions are at the level of items not conditions
  LLMs are not explanatory in the sense of other cogn. models

- case study:
  item-level predictions not supported by the data; condition-level predictions depend on methods of aggregation
  LLMs are "right for the wrong reasons" because the aggregation method that has empirical support is based on an item-level strategy that is ruled out by the data

Summary & conclusions

Benchmark testing

accuracy for gold standard

testing at scale

Machine psychology

careful behavioral diagnostics

human-like behavior

Machine psychology: recommended best practices

- measurement variability
  - prompt-sensitivity
    - formulate different instructions
    - alter order of lists etc.
    - create various instances of the same condition
  - aggregation method
- known LLM biases:
  - token frequency
  - biases towards attending most to beginning and end of prompt (Zhao et al. 2021)
- clear research question / hypothesis
  - what kind of claim to I want to make about which kind of model (Lampinen 2023)
    - dumb prompting vs best possible prompting strategy
    - single-prompt LLM or complex prompting w/ CoT …
- severe testing, not confirmation seeking
  - design to challenge the research hypothesis, not confirm it
  - justify the dependent variable / test measure
  - comparison of model behavior under different types of measures
- preregister analyses
- include statistical tests for main (preregistered) hypotheses
- no overselling of results
- training contamination (Hagendorff 2023)
  - make new material, don't reuse existing ones

cf., Hagendorf (2023)