Panorama of LM evaluations

Clémentine Fourrier

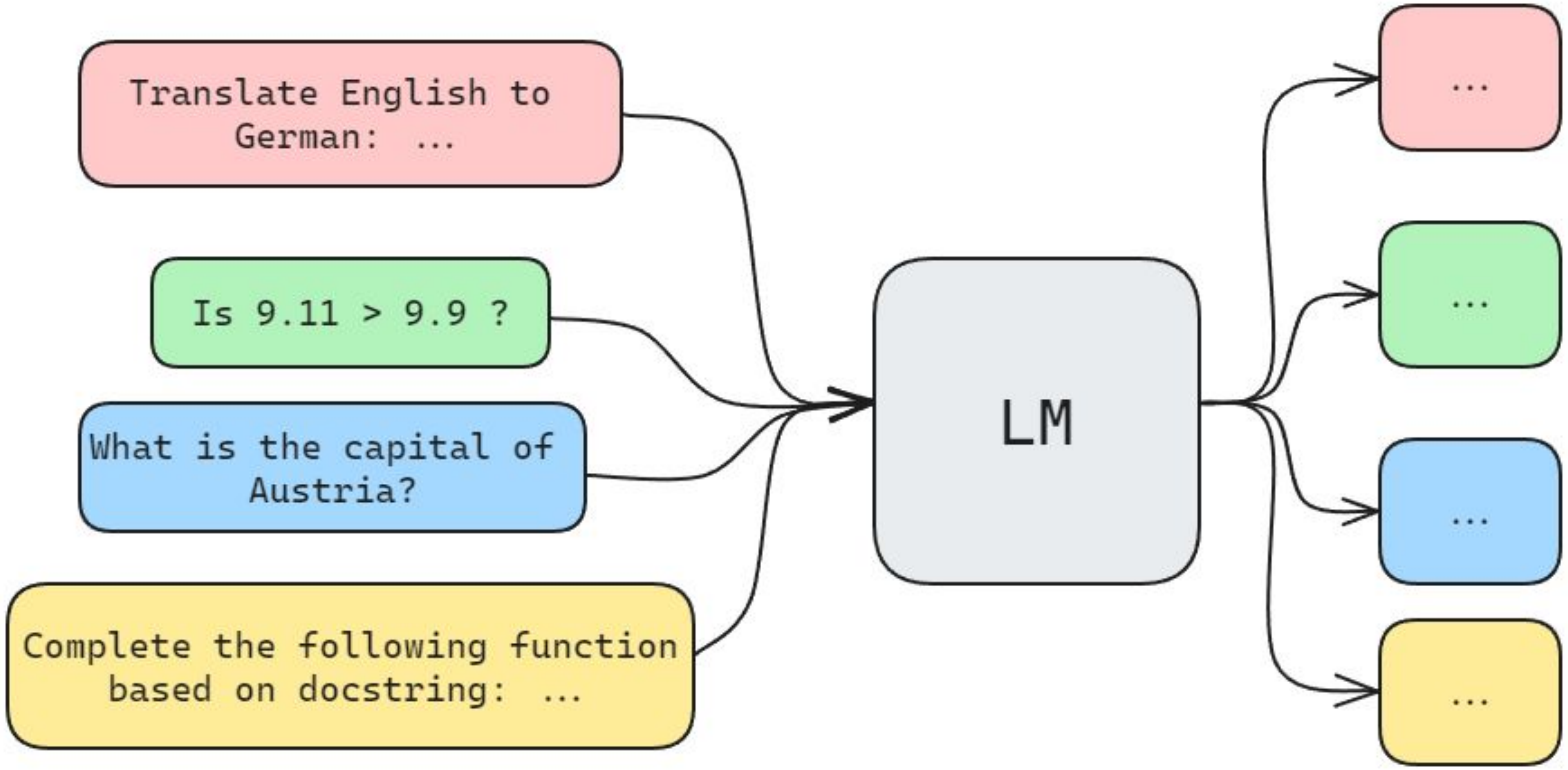
Hugging Face

clefourrier

Spring 2025

Introduction

Language models - capabilities



Why is evaluation important?

Model builders

- best training method
- non-regression
- risks/costs

Users

- best model for X
- hype vs trust

Field

- capabilities
- direction

How to evaluate
Automatic benchmarks

How do you evaluate a language model automatically?

Input from a
dataset
(e.g MMLU)

→

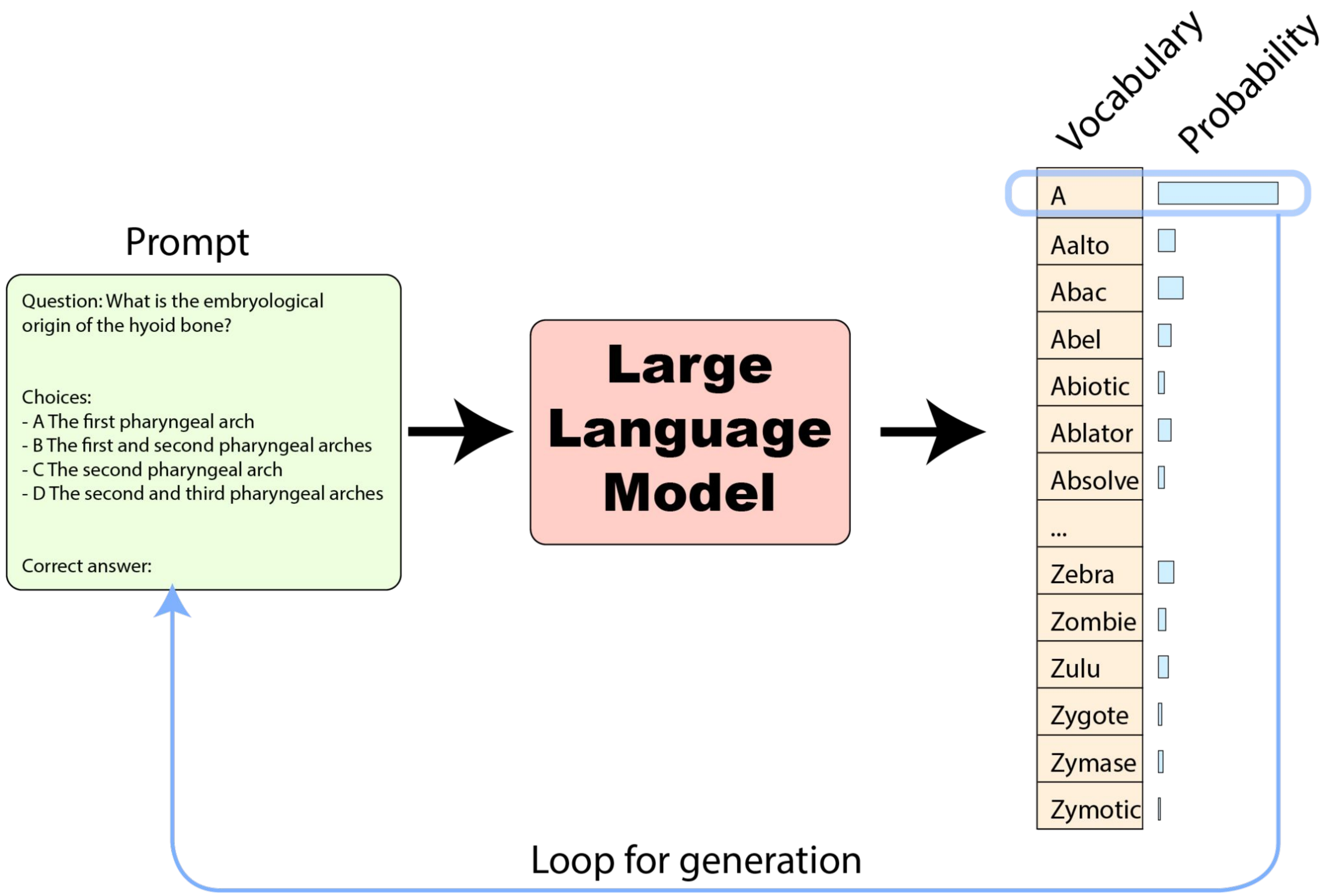
Model
generates a prediction
(e.g words, probabilities)

→

Score the prediction
with a metric
(e.g accuracy, exact match,
BLEU, ROUGE, …)

How do you evaluate a language model automatically?

Input from a **dataset** (e.g MMLU) → Model **generates a prediction** (e.g words, probabilities) → Score the prediction **with a metric** (e.g accuracy, exact match, BLEU, ROUGE, …)

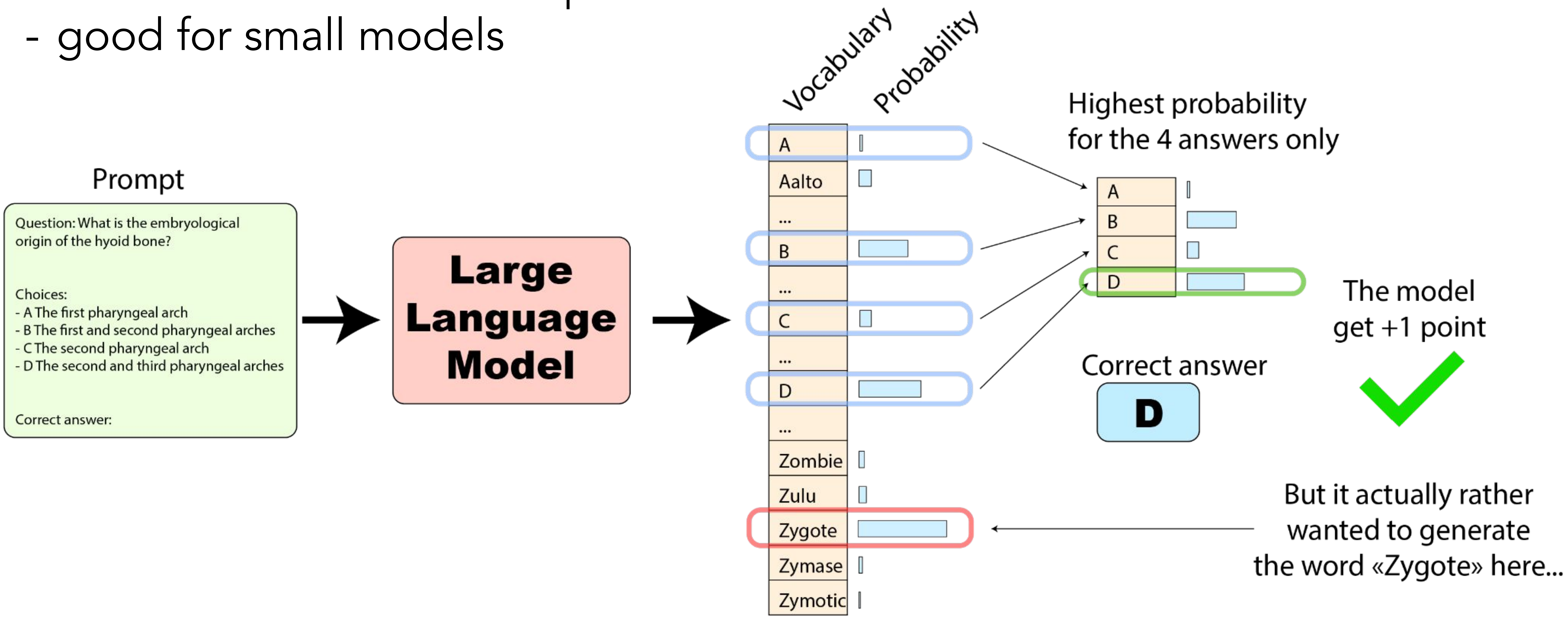


https://huggingface.co/blog/open-llm-leaderboard-mmlu

Clémentine Fourrier

2 ways to get a prediction

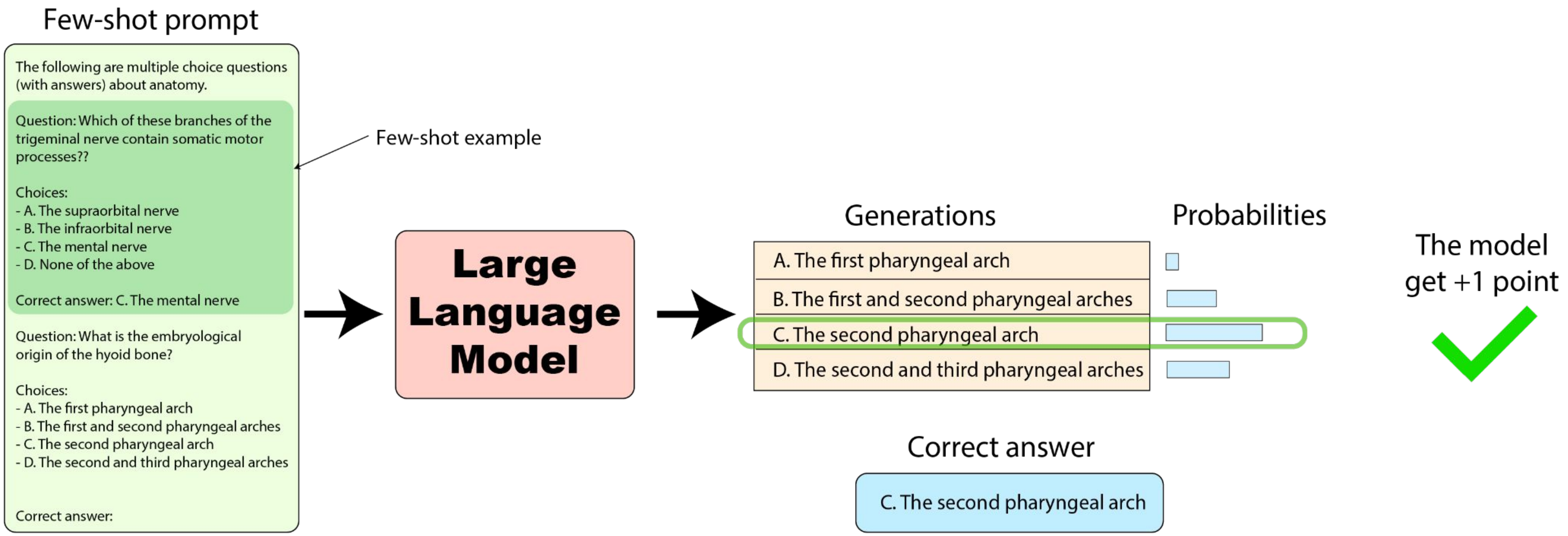
Probabilities based evals:
- constrain the evaluation space
- good for small models



https://huggingface.co/blog/open-llm-leaderboard-mmlu

2 ways to get a prediction

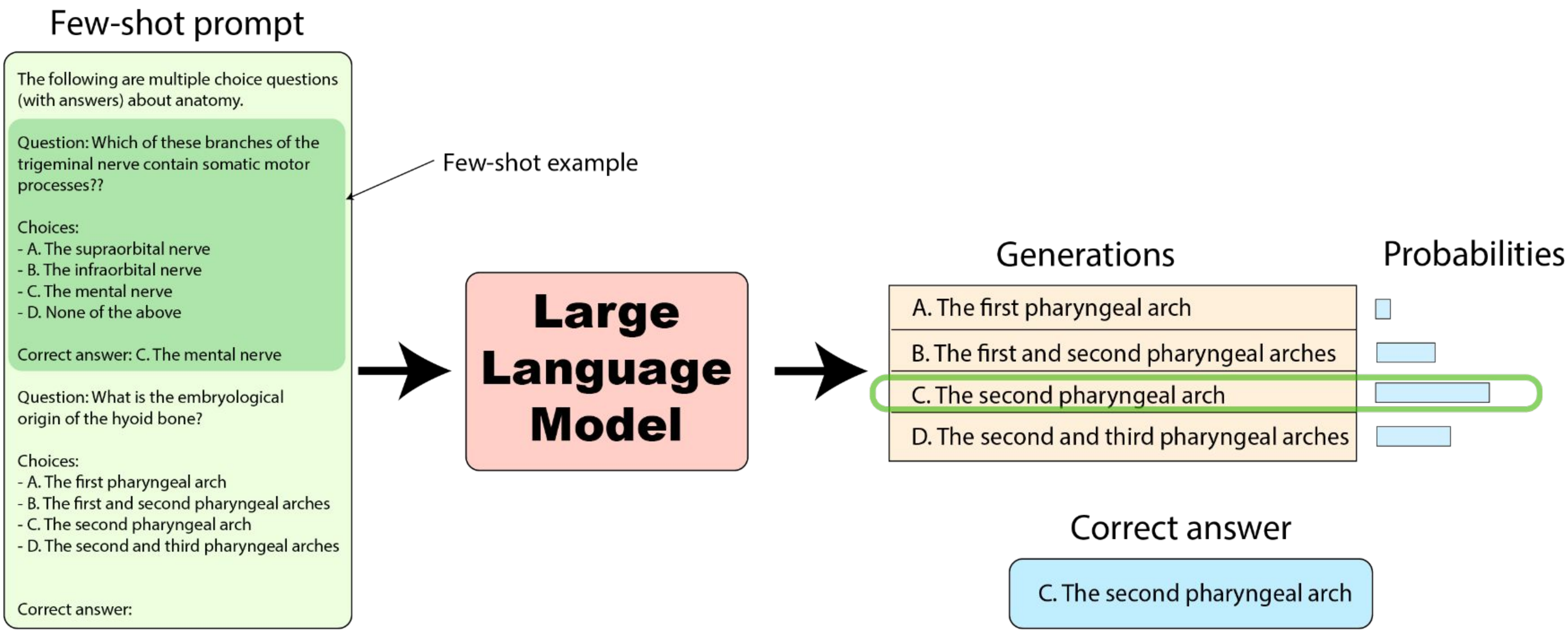
Generation based evals:
- closer to real world use cases
- harder to score



https://huggingface.co/blog/open-llm-leaderboard-mmlu

Scoring a free form prediction

In context learning/providing examples/few-shot



https://huggingface.co/blog/open-llm-leaderboard-mmlu

Scoring a free form prediction

Prompt for a format

System prompt: You are a general AI assistant. I will ask you a question. Report your thoughts, and finish your answer with the following template: FINAL ANSWER: [YOUR FINAL ANSWER]. YOUR FINAL ANSWER should be a number OR as few words as possible OR a comma separated list of numbers and/or strings.
If you are asked for a number, don't use comma to write your number neither use units such as $ or percent sign unless specified otherwise.
If you are asked for a string, don't use articles, neither abbreviations (e.g. for cities), and write the digits in plain text unless specified otherwise.
If you are asked for a comma separated list, apply the above rules depending of whether the element to be put in the list is a number or a string.

GAIA Question: The attached Excel file contains the sales of menu items for a local fast-food chain. What were the total sales that the chain made from food (not including drinks)? Express your answer in USD with two decimal places.

XLSX

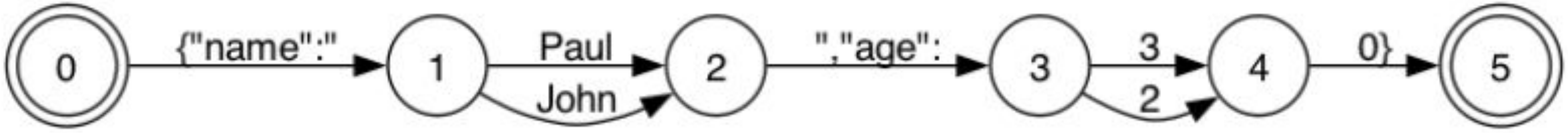
`uploaded.xlsx`

GAIA: *https://arxiv.org/pdf/2311.12983*

Scoring a free form prediction

Constraining the output with structured text generation

```
{
    "name": "John"|"Paul",
    "age": 20|30
}
```



Character JSON regex FSM Optimized

https://blog.dottxt.co/coalescence.html

Scoring a free form prediction

Improving answer extraction with smart parsing

Example: MATH dataset

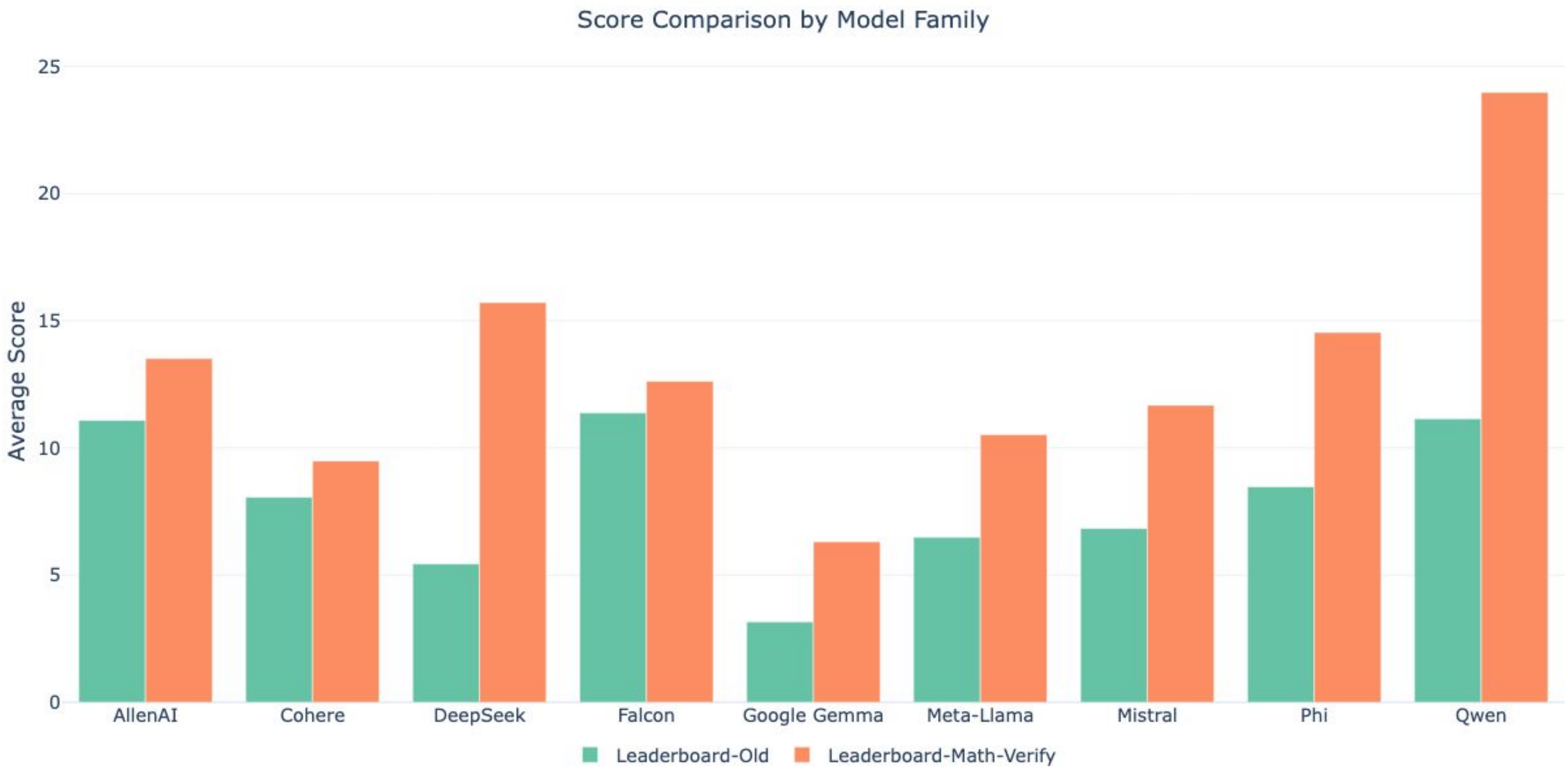
Answer should follow:
"Final answer is [ANSWER].
I hope it is correct."

📄 Example	❗ Issue	✅ Math-Verify
The final answer is $2x + 4y + z - 19 = 0$. I hope it is correct.	Partial parse of parametric eq	Eq(2x + 4y + z - 19, 0)
(23)	Failed extraction due to latex borders	23
((- \infty, -14) \cup (-3, \infty)).	Failed extraction due to interval	Union(Interval.open(-oo, -14), Interval.open(-3, oo))
100%	Failed extraction due to invalid symbol	1
\begin{pmatrix}\frac{1}{50}&\frac{7}{50}\frac{7}{50}&\frac{49}{50}\end{pmatrix}	Failed extraction due to Matrix	Matrix([[1/50, 7/50], [7/50, 49/50]])

https://huggingface.co/blog/math_verify_leaderboard

Scoring a free form prediction

Improving answer extraction with smart parsing

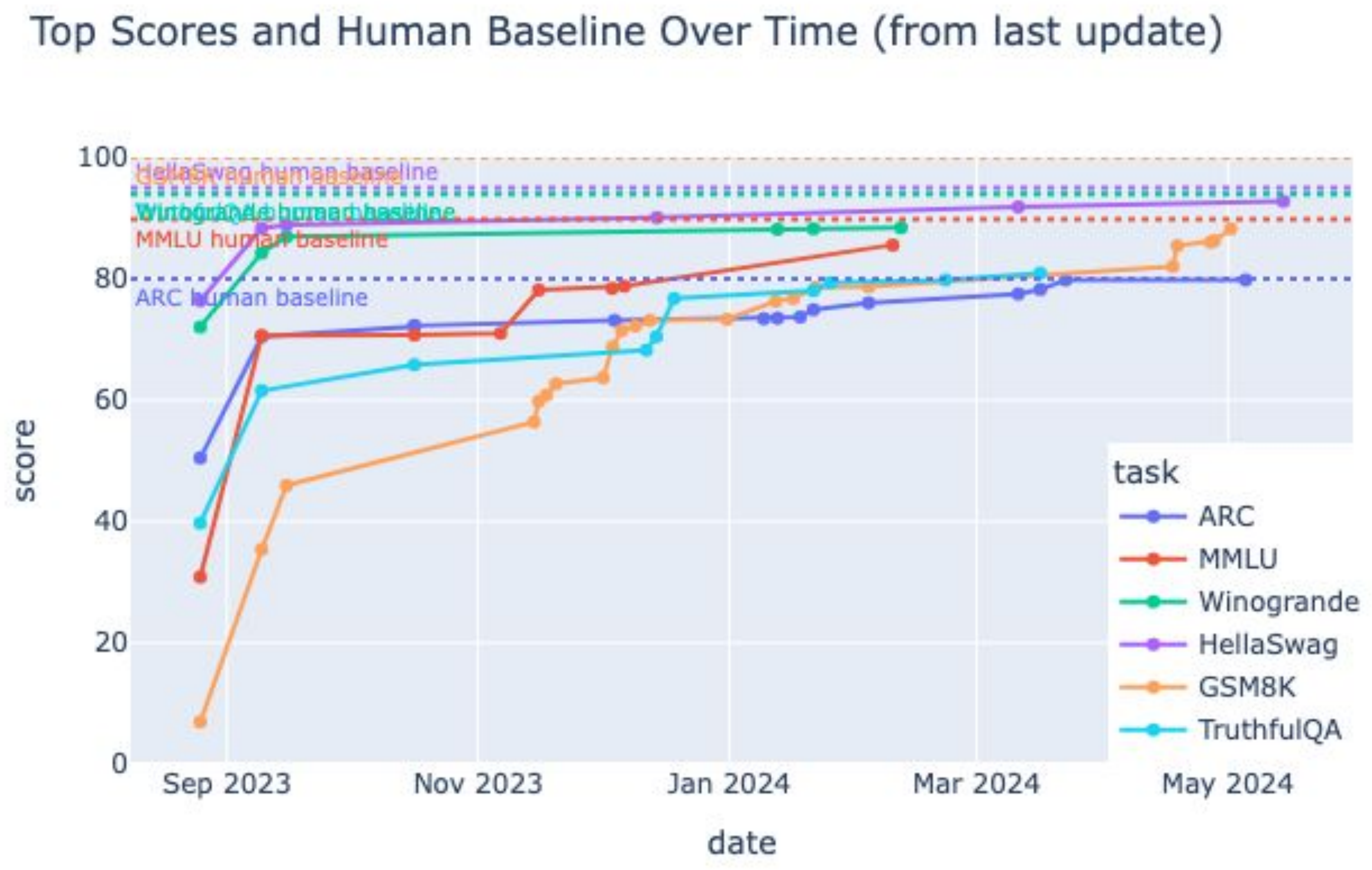


https://huggingface.co/blog/math_verify_leaderboard

How do you evaluate a language model automatically?

Input from a dataset (e.g MMLU) → Model generates a prediction (e.g words, probabilities) → Score the prediction with a metric (e.g accuracy, exact match, BLEU, ROUGE, …)

Should:
- Reflect your use case
- Be unseen :/
- Be unsaturated



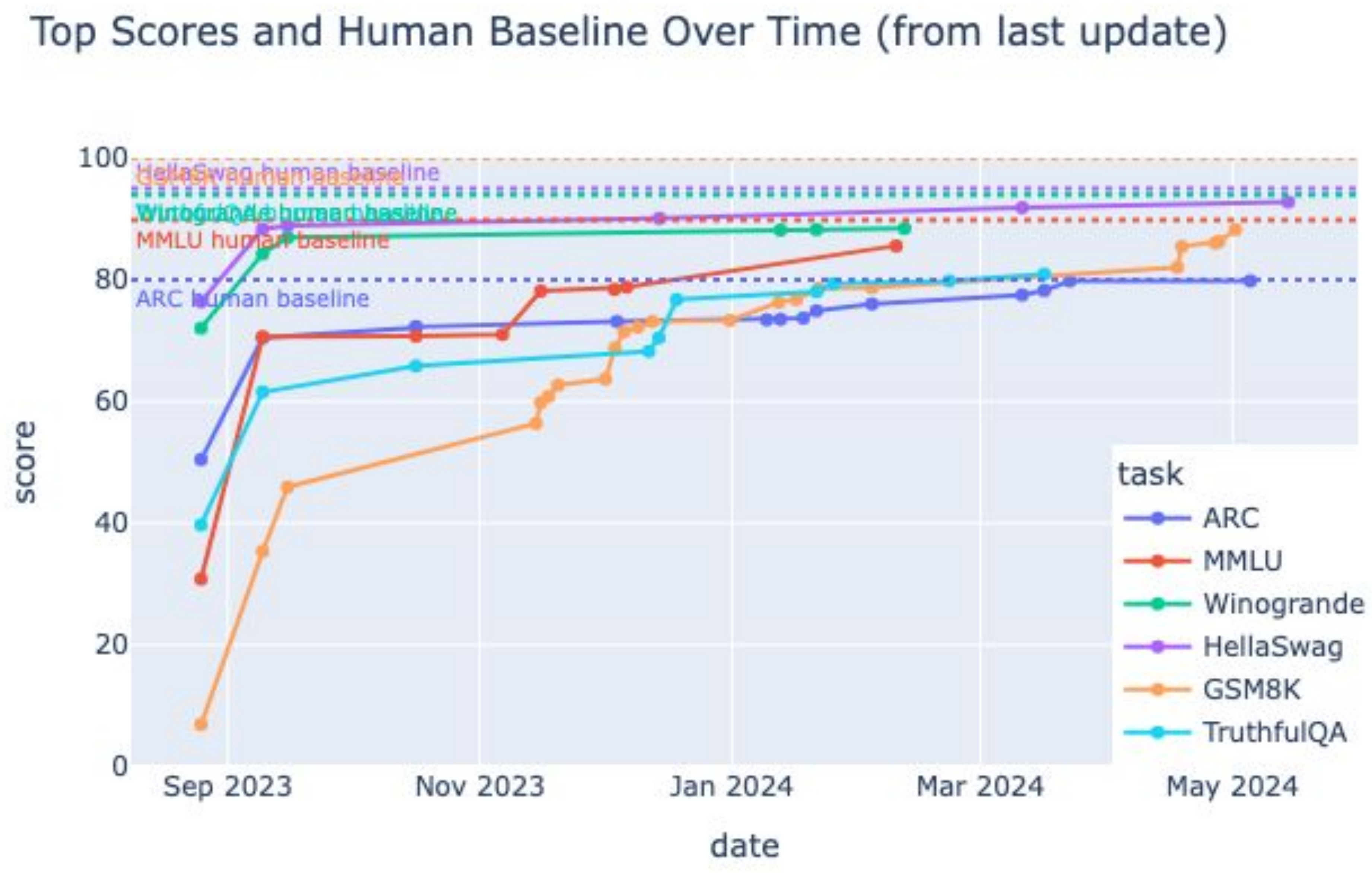
https://github.com/huggingface/evaluation-guidebook/blob/main/contents/automated-benchmarks/some-evaluation-datasets.md
https://huggingface.co/evaluate-metric

How do you evaluate a language model automatically?

Input from a **dataset** (e.g MMLU) → Model **generates a prediction** (e.g words, probabilities) → Score the prediction **with a metric** (e.g accuracy, exact match, BLEU, ROUGE, …)

Should:
- Reflect your use case
- Be unseen :/
- Be unsaturated

Inspect:
- Questions: MMLU -> MMLU-(Redux/Global/Pro)
- Process: Experts > Annotators > MTurkers



https://github.com/huggingface/evaluation-guidebook/blob/main/contents/automated-benchmarks/some-evaluation-datasets.md
https://huggingface.co/evaluate-metric

Clémentine Fourrier

How do you evaluate a language model automatically?

Input from a dataset (e.g MMLU) → Model generates a prediction (e.g words, probabilities) → Score the prediction with a metric (e.g accuracy, exact match)

Pros:
- consistency, reproducibility
- limited cost
- understandability of metrics

Cons:
- hard to evaluate real life use cases
  - chat models - 2022
  - reasoning models - 2025
- contamination

How to evaluate
Automatic benchmarks: Unit testing

Unit testing

Input from a
dataset
(e.g HumanEval, IFEval)

→

Model
generates a prediction
(words)

→

Prediction must
satisfy a condition
(e.g pass a test)

Unit testing for language

Input from a
dataset
(e.g HumanEval, IFEval)

Model
generates a prediction
(words)

Prediction must
satisfy a condition
(e.g pass a test)

Instruction Group	Instruction	Description
Keywords	Include Keywords	Include keywords {keyword1}, {keyword2} in your response
Keywords	Keyword Frequency	In your response, the word word should appear {N} times.
Keywords	Forbidden Words	Do not include keywords {forbidden words} in the response.
Keywords	Letter Frequency	In your response, the letter {letter} should appear {N} times.
Language	Response Language	Your ENTIRE response should be in {language}, no other language is allowed.
Length Constraints	Number Paragraphs	Your response should contain {N} paragraphs. You separate paragraphs using the markdown divider: * * *
Length Constraints	Number Words	Answer with at least / around / at most {N} words.
Length Constraints	Number Sentences	Answer with at least / around / at most {N} sentences.
Length Constraints	Number Paragraphs + First Word in i-th Paragraph	There should be {N} paragraphs. Paragraphs and only paragraphs are separated with each other by two line breaks. The {i}-th paragraph must start with word {first_word}.
Detectable Content	Postscript	At the end of your response, please explicitly add a postscript

Used for code models:
- passing unit tests

IFEval:
- unit tests for language

How to evaluate
Human evaluations

How do you evaluate a language model with humans?

Input from a human (sometimes from a dataset) → Model generates a prediction (words) → Score the prediction with a human (e.g grade, preference)

How do you evaluate a language model with humans?

Vibe check

- getting a feel
- testing on your use case

Arena

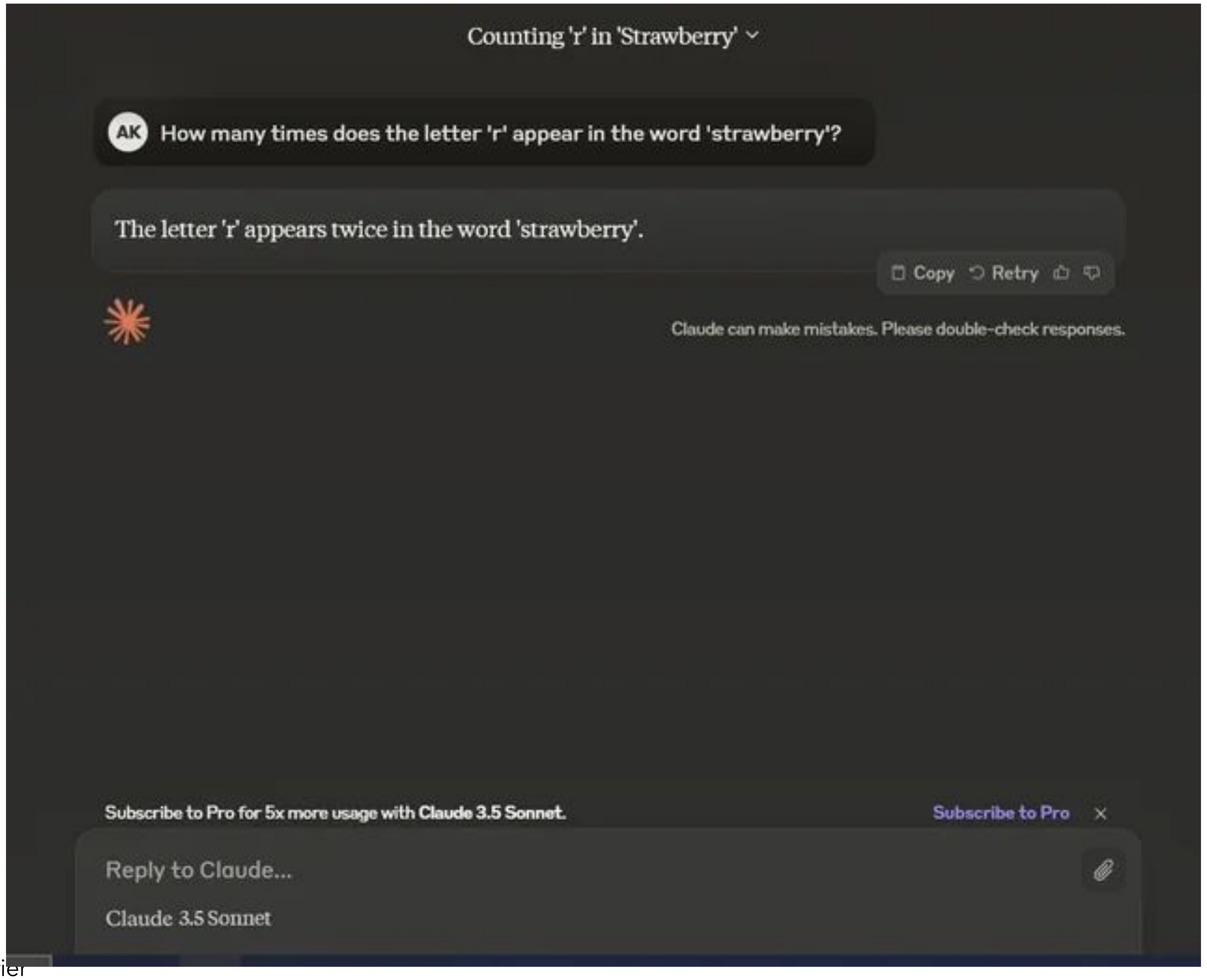
- vibe-checks at scale
- edge case discovery

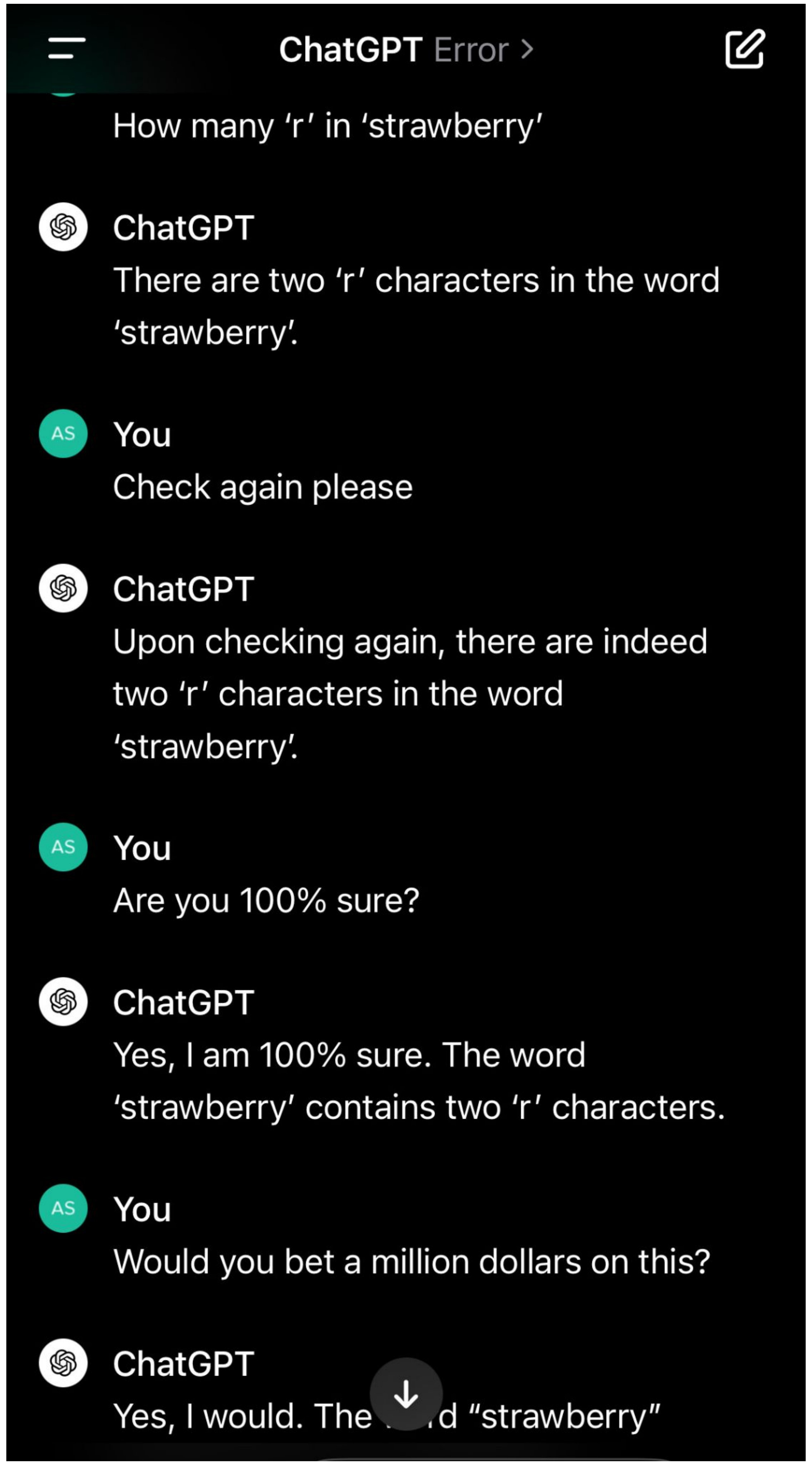
Systematically

- strict guidelines
- paid annotators

How do you evaluate a language model with humans?

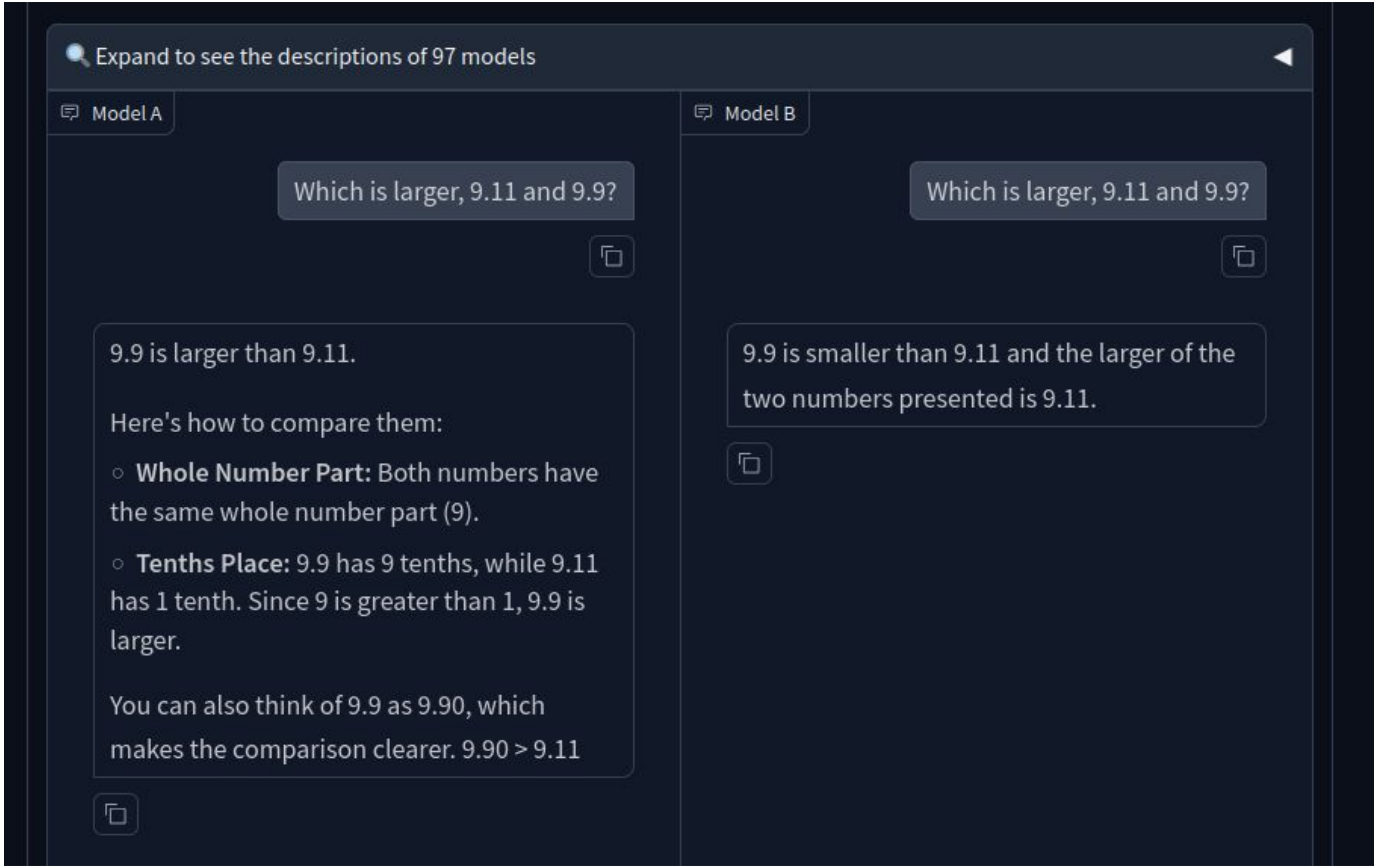
How many r in strawberry?
9.11 and 9.9, which is larger?
Draw me a unicorn in tikz/latex/…





Posts on X

How do you evaluate a language model with humans? ⚔️



https://lmarena.ai/

How do you evaluate a language model with humans? 🧐 ⚔️

Towards Understanding Sycophancy in Language Models

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, Ethan Perez

Human feedback is commonly utilized to finetune AI assistants. encourage model responses that match user beliefs over truthfu sycophancy. We investigate the prevalence of sycophancy in mo use of human feedback, and the potential role of human prefere first demonstrate that five state-of-the-art AI assistants consister varied free-form text-generation tasks. To understand if human p

Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation

Nitesh Goyal, Ian Kivlichan, Rachel Rosen, Lucy Vasserman

Machine learning models are commonly used to detect toxicity in online conversations. These models are trained on datasets annotated by human raters. We explore how raters' self-described identities impact how they annotate toxicity in online comments. We first define the concept of specialized rater pools: rater pools formed based on raters' self-described identities, rather than at random. We formed

Human Feedback is not Gold Standard

Tom Hosking, Phil Blunsom, Max Bartolo

Human feedback has become the de facto standard for evaluating the performance of Large Language Models, and is increasingly being used as a training objective. However, it is not clear which properties of a generated output this single `preference' score captures. We hypothesise that preference scores are subjective and open to undesirable biases. We critically analyse the use of human feedback for both training and evaluation, to verify whether it fully captures a range of crucial error criteria. We find that while preference scores have fairly good coverage, they under-represent important aspects like factuality. We further hypothesise that both preference scores and error annotation may be affected by

- biased (first impression, assertiveness, self preference, …)
- easy to game
- subjective/unreproducible
- not too costly

How do you evaluate a language model with humans?

Annotation Process
- Iterative Annotation
- Careful Data Selection
- Annotation Scheme
- Guideline Design
- Pilot Study
- Validation Step

Annotators
- Workforce Selection
- Qualification Test
- Annotator Training
- Annotator Debriefing
- Monetary Incentive

Quality Estimation
- Error Rate
- Control Questions
- Agreement

Quality Improvement
- Correction
- Updating Guidelines
- Filtering
- Annotator Feedback
- Annotator Deboarding

Adjudication
- Manual Curation
- Majority Voting
- Probabilistic Aggregation

Figure 1
Quality Management methods discussed in this work. We categorize methods into annotation process, annotator management, quality estimation, quality improvement, and adjudication.

Keep in mind
- simple is better
- remove unnecessary info/simplify to reduce bias
- independent work of annotators
- consistent guidelines
- consider hybrid annotations

- costly
- can fit a specific use case
- but beware of bias still

https://aclanthology.org/2024.cl-3.1/

How to evaluate Model as a judge

How do you evaluate a language model with a model?

Input from a dataset → Model generates a prediction (words) → Score the prediction with a model

Requirements:
- dataset
- precise prompt
- good enough judge model

How do you evaluate a language model with a model?

Input from a dataset → Model generates a prediction (words) → Score the prediction with a model

Requirements:
- dataset
- precise prompt
- good enough judge model

Pros:
- scalable
- cheaper
- reproducible if you use OSS

Cons:
- filled with hard to debug hidden biases
- need to evaluate your evaluator

How do you evaluate a language model with a model?

Bias, bias everywhere

LLM Evaluators Recognize and Favor Their Own Generations

Arjun Panickssery, Samuel R. Bowman, Shi Feng

Self-evaluation using large language models (LLMs) has proven valuable not only in benchmarking but also methods like reward modeling, constitutional AI, and self-refinement. But new biases are introduced due to the same LLM acting as both the evaluator and the evaluatee. One such bias is self-preference, where an LLM evaluator scores its own outputs higher than others' while human annotato

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena

Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
ıeng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica

e model (LLM) based chat assistants is challenging due to their broad
quacy of existing benchmarks in measuring human preferences. To address
ong LLMs as judges to evaluate these models on more open-ended questions.
nd limitations of LLM-as-a-judge, including position, verbosity, and self-
well as limited reasoning ability, and propose solutions to mitigate some of
agreement between LLM judges and human preferences by introducing two

Finding Blind Spots in Evaluator LLMs with Interpret Checklists

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, Mite

Large Language Models (LLMs) are increasingly relied upon to evaluate text outputs of other LLMs, thereby influencing leaderboards and development decisions. However, concerns persist over the accuracy of these assessments and the potential for misleading conclusions. In this work, we

Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators

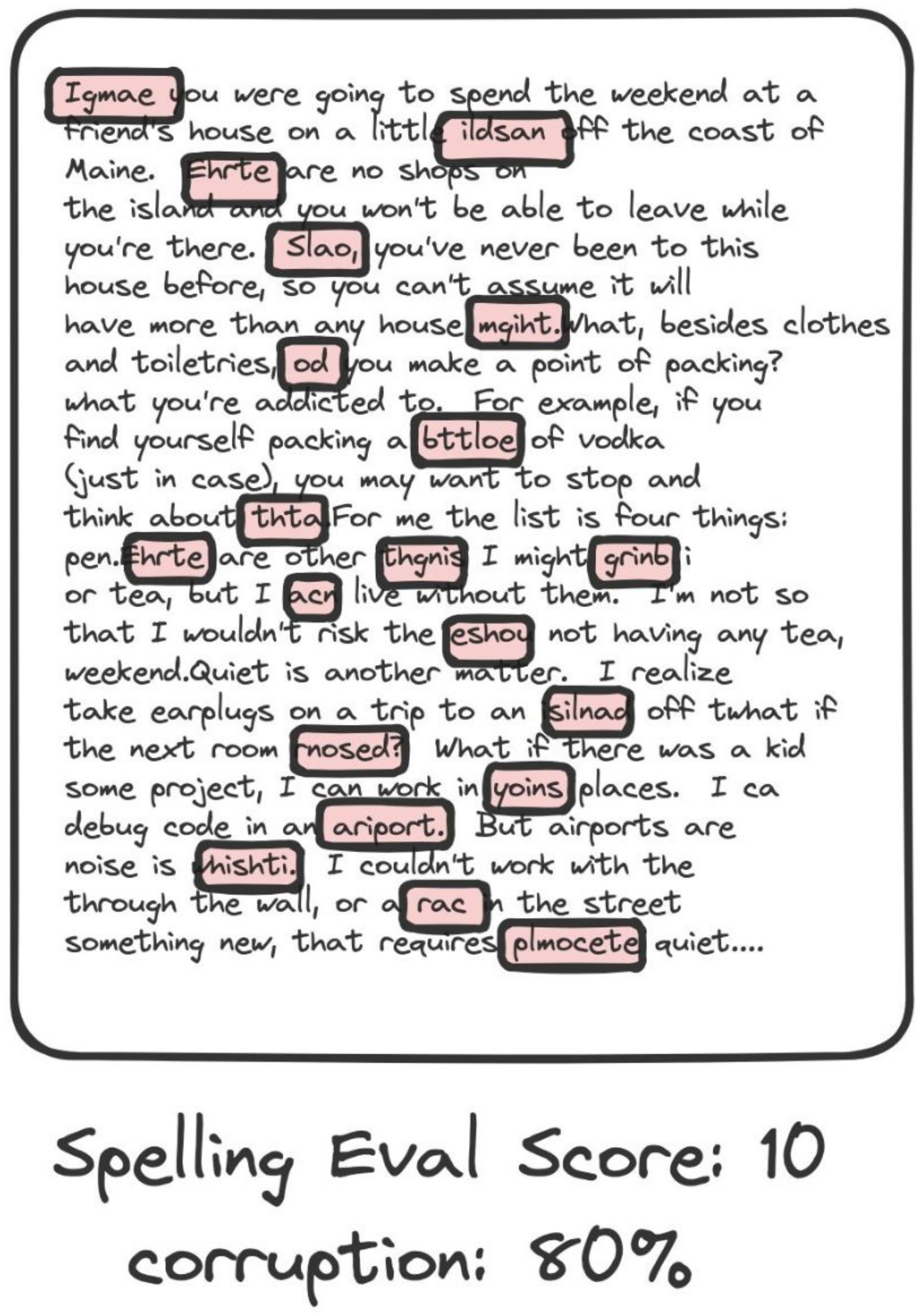
Yann Dubois, Balázs Galambosi, Percy Liang, Tatsunori B. Hashimoto

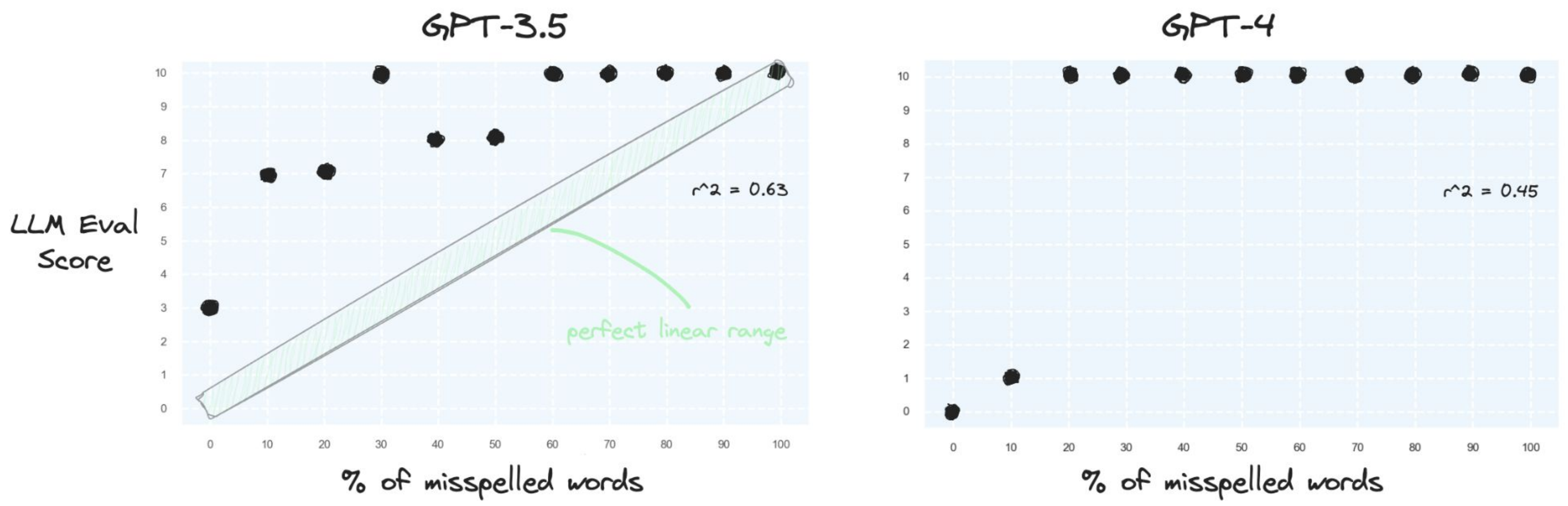
LLM-based auto-annotators have become a key component of the LLM development process due to their cost-effectiveness and scalability compared to human-based evaluation. However, these auto-annotators can introduce complex biases that are hard to remove. Even simple, known confounders such as preference for longer outputs remain in existing automated evaluation metrics. We propose a

- Self preference bias
- Position bias
- Verbosity bias
- Format bias
- Lack of internal consistency

How do you evaluate a language model with a model?

Bias, bias everywhere (blindness to perturbation, inability to score on a scale)





Scoring Template

- Score 0 indicates the document is free of grammatical and spelling errors.
- Score 2 signifies that 20% of the words contain errors.
- Score 5 indicates that 50% of the words are erroneous.
- Score 7 reflects 70% error prevalence.
- Score 10 means that every word in the document has grammatical errors.

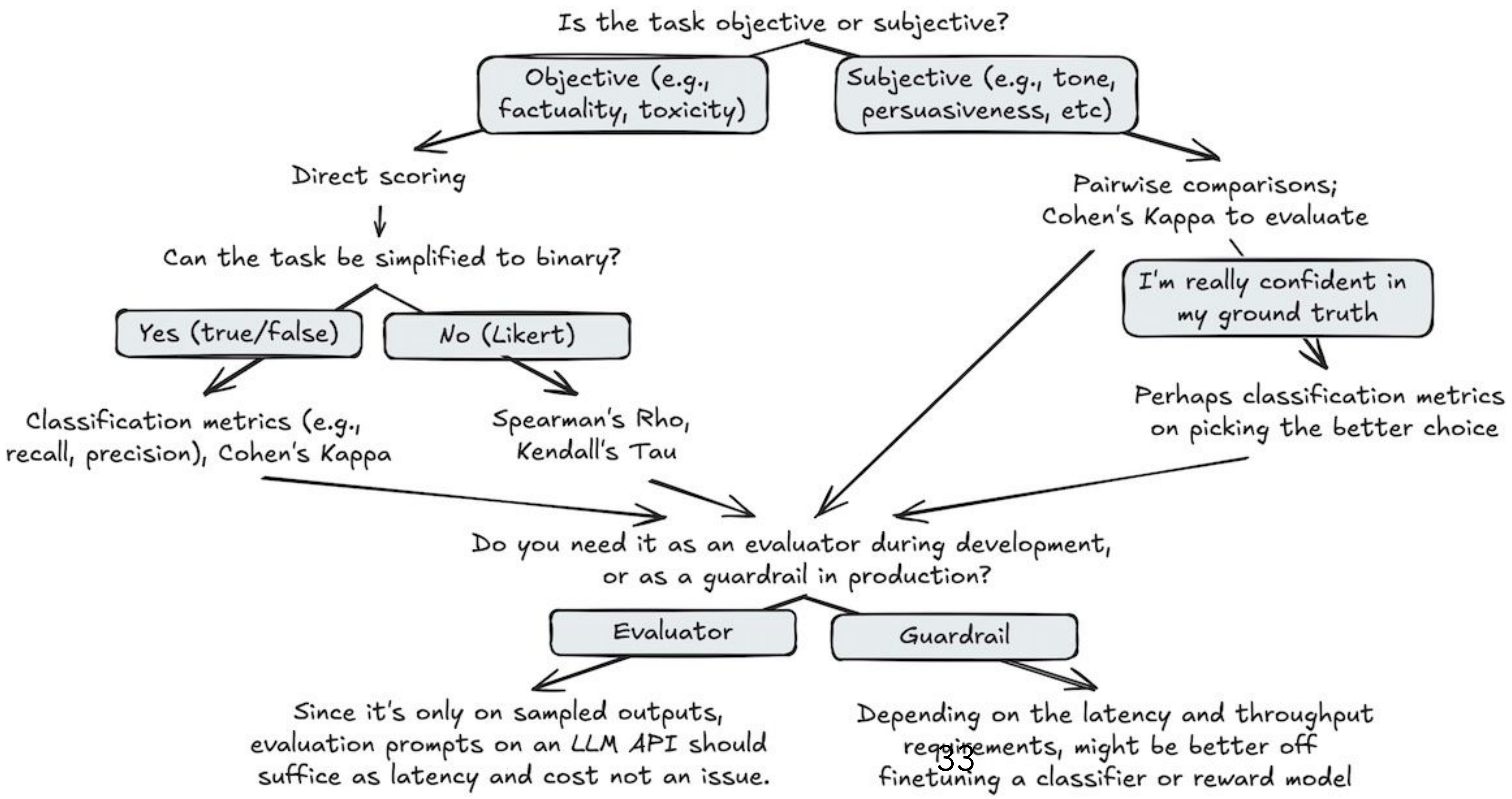
https://github.com/LeonEricsson/llmjudge/blob/main/README.md

https://x.com/aparnadhinak/status/1748368364395721128/photo/1

How do you evaluate a language model with a model?

- Lack of internal consistency -> judge multiple prompting
- Self preference -> using a jury
- Inconsistent score ranges -> asking to justify the score, providing the scale in the prompt
- Position bias -> switching positions randomly
- Verbosity bias -> normalize the score with the length

…



https://eugeneyan.com/assets/llm-eval-tree.jpg

Evaluation in practice

Why is evaluation important?

Model builders

- best training method
- non-regression
- risks/costs

Users

- hype vs trust
- best model for X

Field

- capabilities
- direction

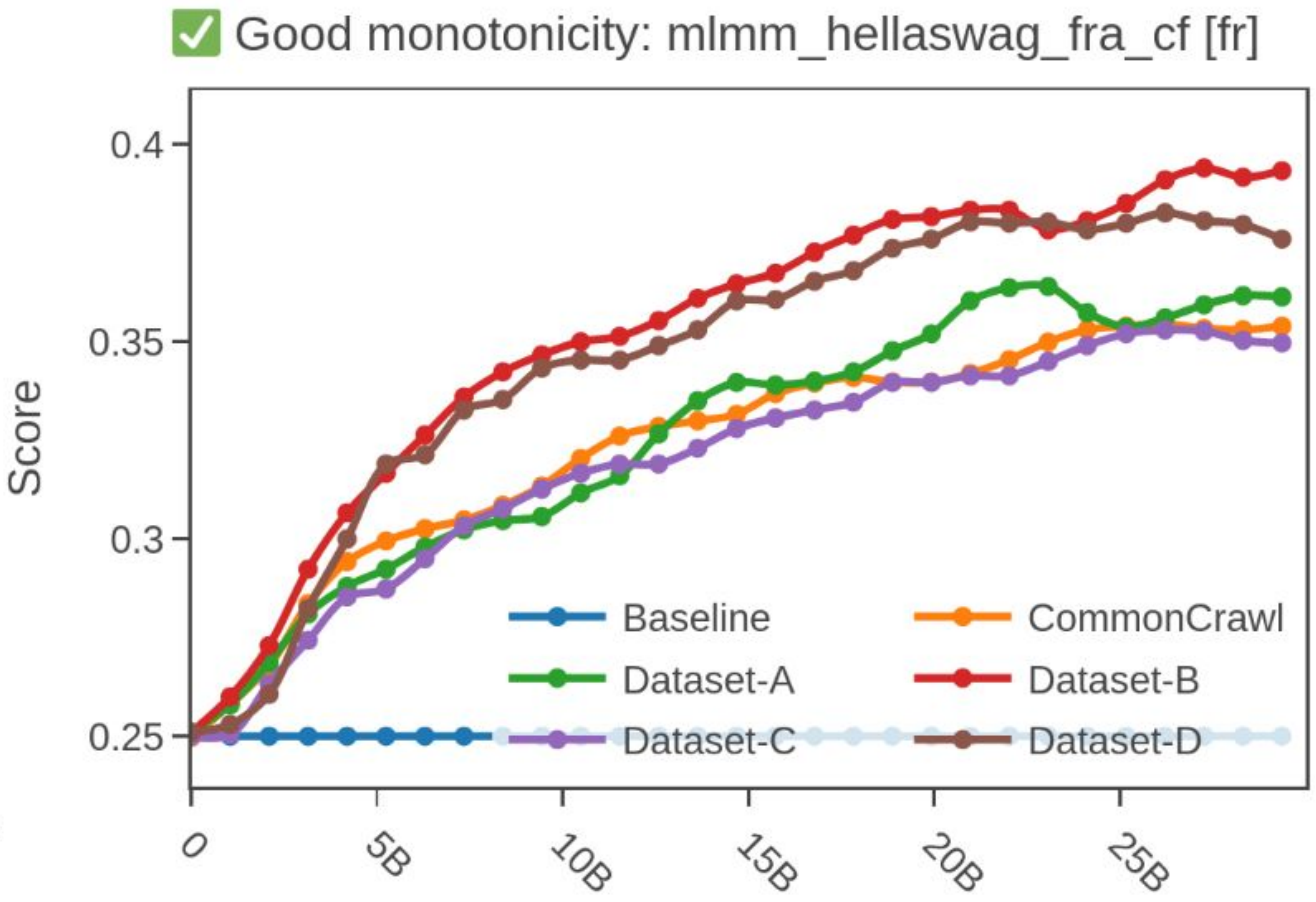
Evaluation in practice
Finding high-signal evaluation for training

Slides from this section are by Guilherme Penedo, of the FineWeb team at HF

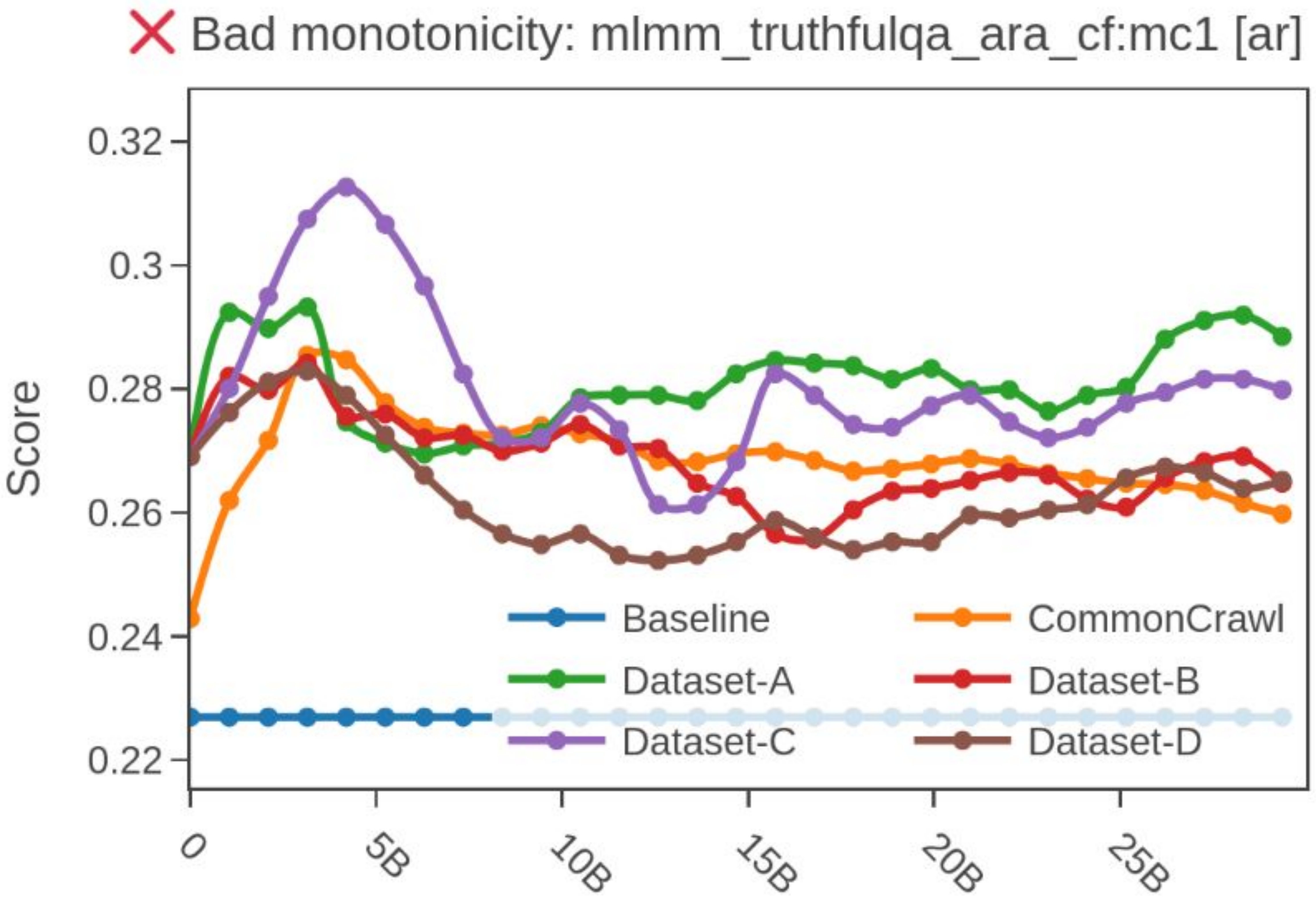
High-signal: monotonicity

Rationale: We should see learning as training progresses

Measure: *Spearman rank correlation* between steps and score



Monotonicity: 0.98

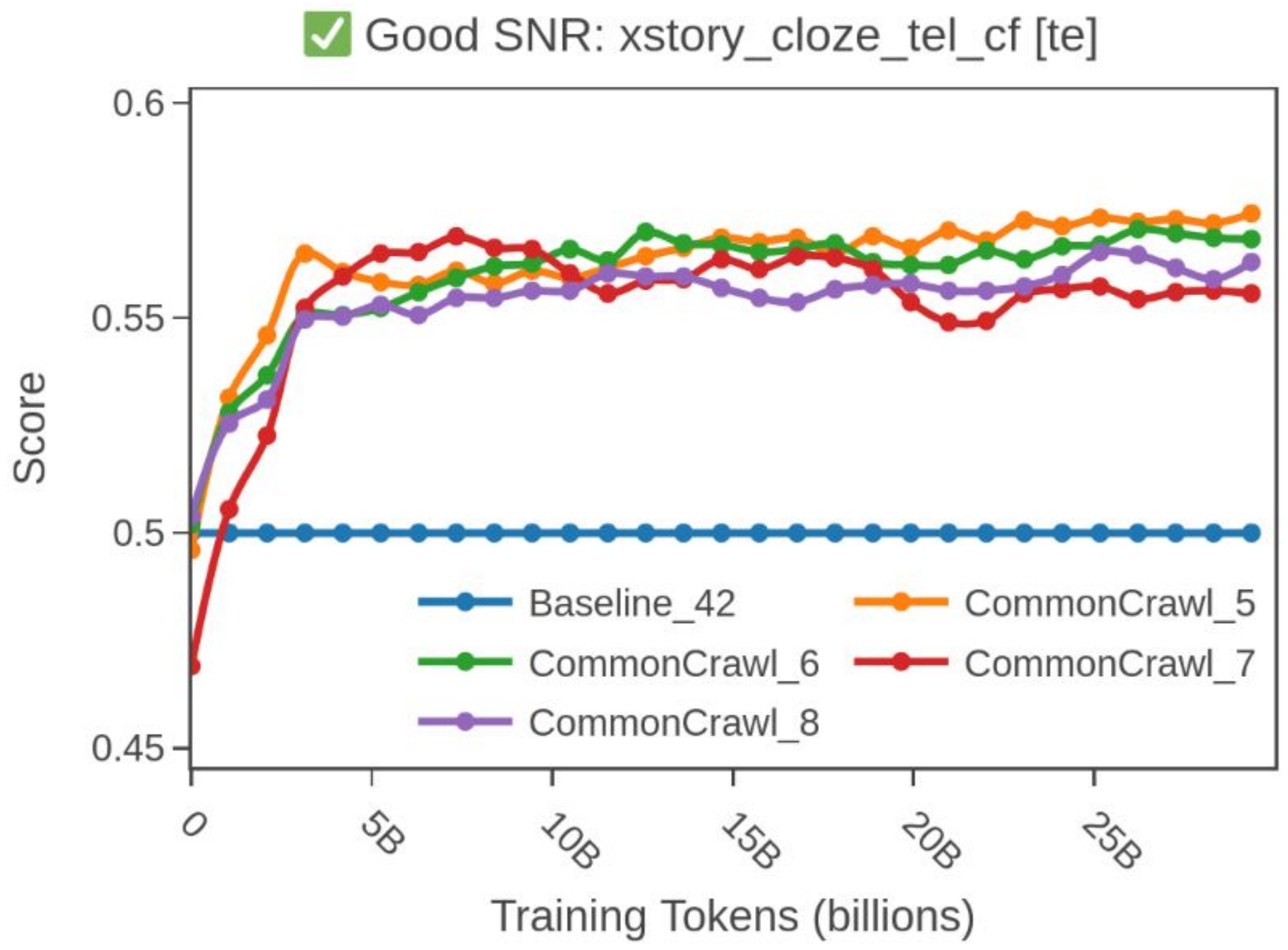


Monotonicity: -0.26

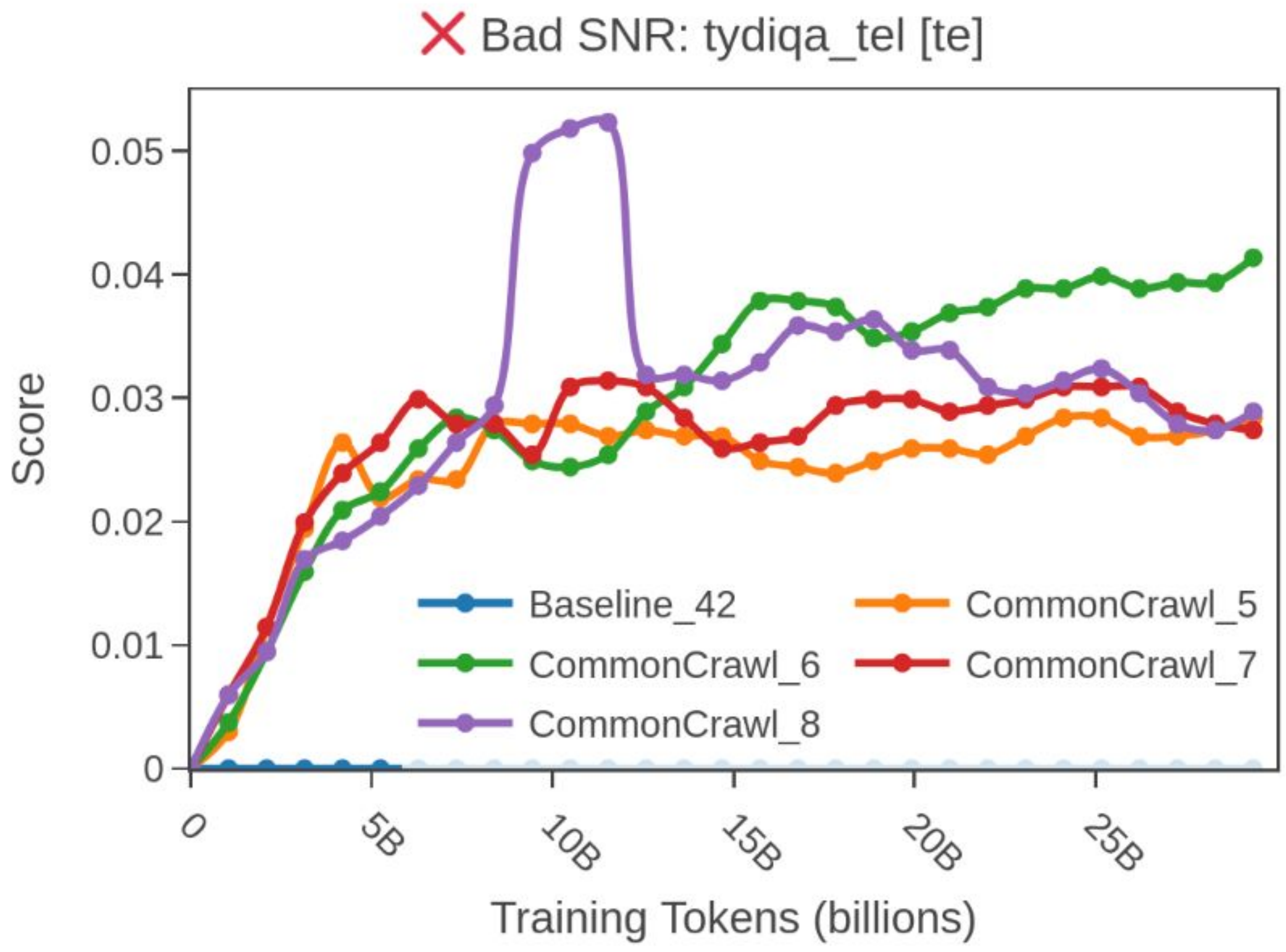
High-signal: low noise

Rationale: Score differences should not be caused by evaluation noise

Measure: *SNR = (avg score / std_dev);* with std_dev coming from diff seeds of “noisy” data



Signal-to-Noise: 87.87

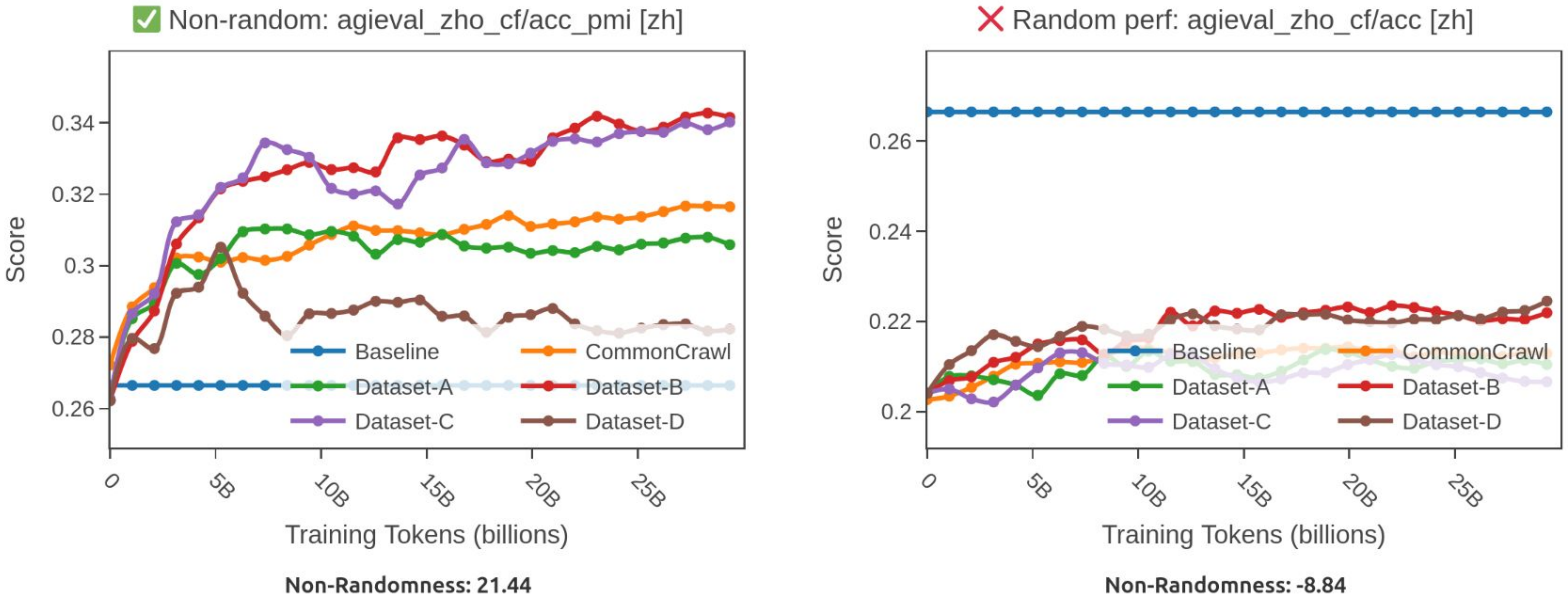


Signal-to-Noise: 6.40

High-signal: above random

Rationale: Can not conclude anything if the model has random performance [for pretraining ablations!]

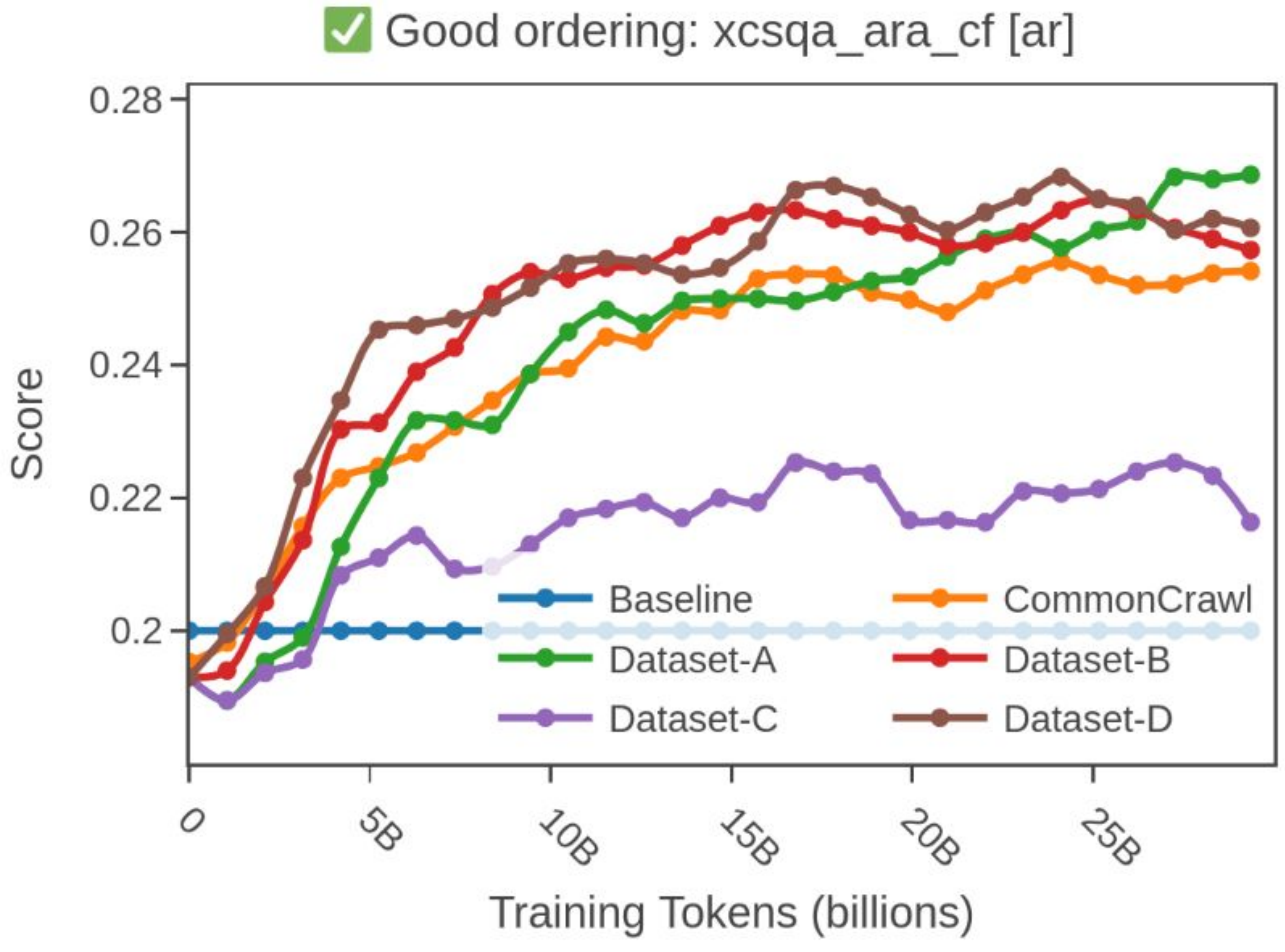
Measure: *Max distance to RB in std_dev;* with std_dev coming from diff seeds of "noisy" data



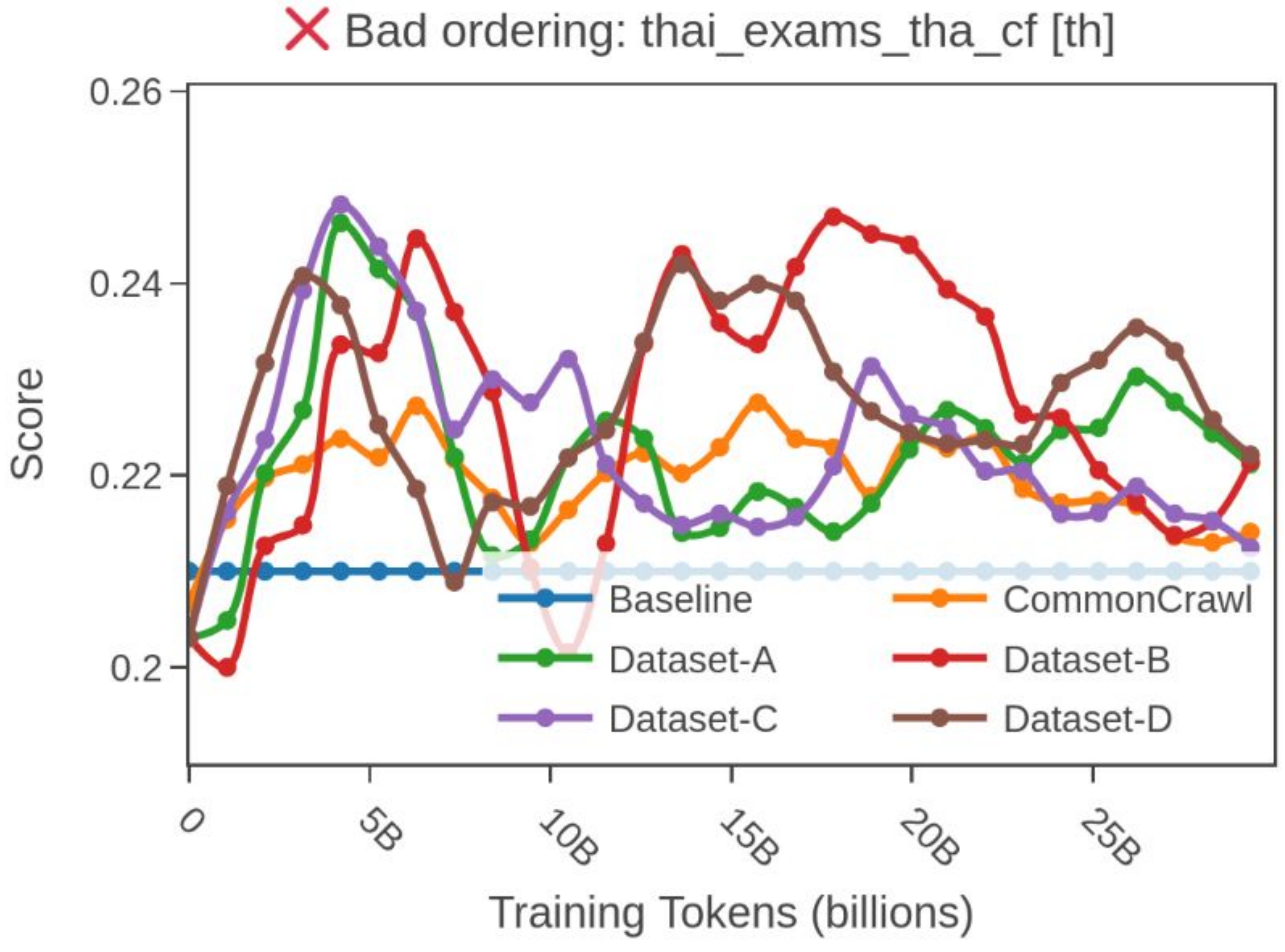
High-signal: ordering consistency

Rationale: We want to generalize to larger scales, pre-condition for that is stable ordering at the experiment scale

Measure: *Kendall-tau for every consecutive step pair*



Ordering Consistency: 0.83



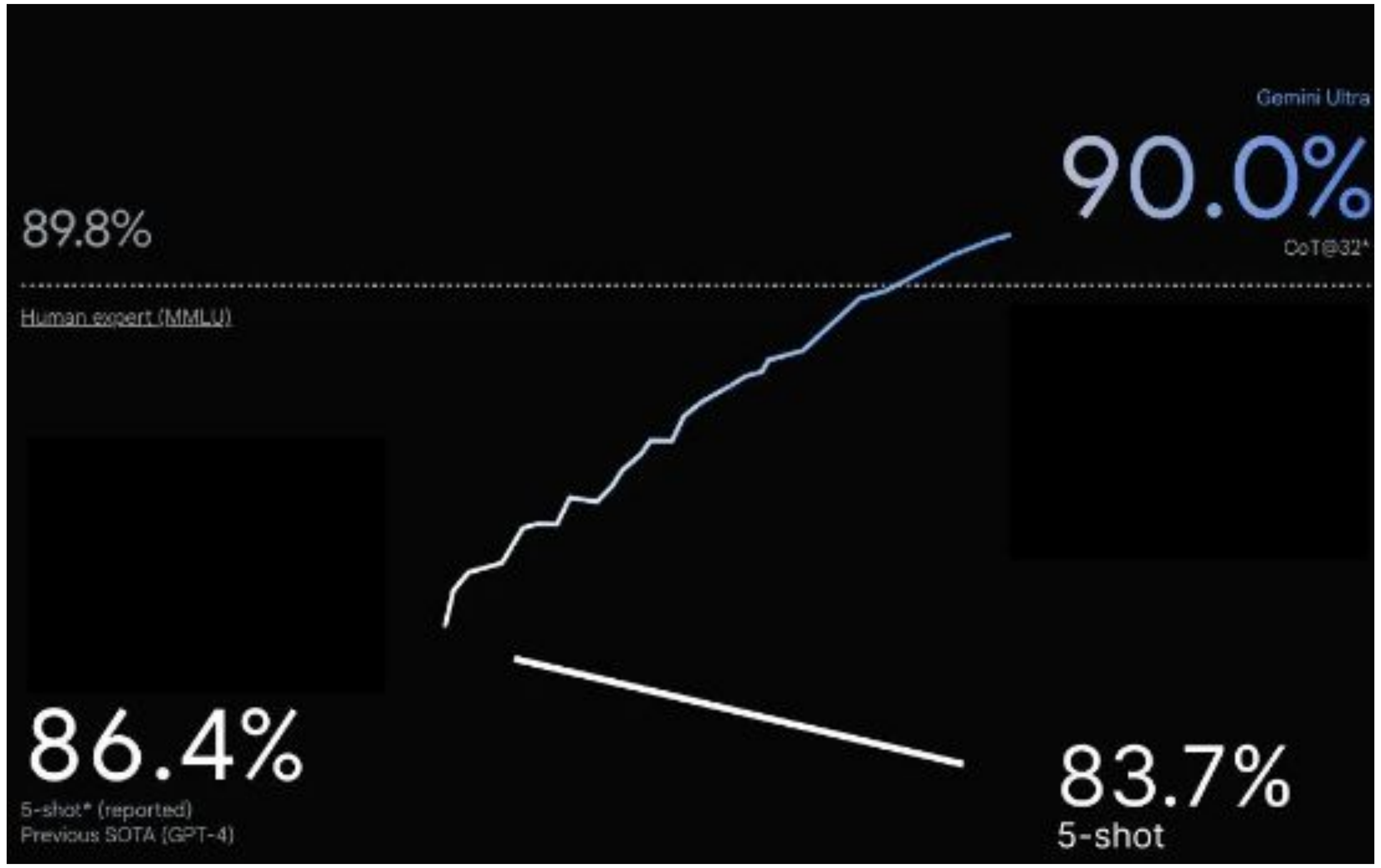
Ordering Consistency: 0.69

Evaluation in practice
Cutting through the hype, or why you can't reproduce scores of the latest release

Task specific issues

Not using the same metric
- probability vs generation metric
- normalisation of outputs (numbers, punctuation, …)
- actually reporting different metrics

```yaml
metric_list:
  - metric: exact_match
    aggregation: mean
    higher_is_better: true
    ignore_punctuation: true
    ignore_case: true
```



https://github.com/EleutherAI/lm-evaluation-harness/blob/main/lm_eval/tasks/mmlu/generative/_default_template_yaml

"Corrected" gemini announcement, PSchmid, X

Task specific issues

Not using the same parameters
- for generation
    - temperature
    - termination management (token, length)
- for the model
    - randomness seeds
    - batch size
    - weight precision

Prompt specific issues

Prompting method and model types: LM > Chat > Reasoning models

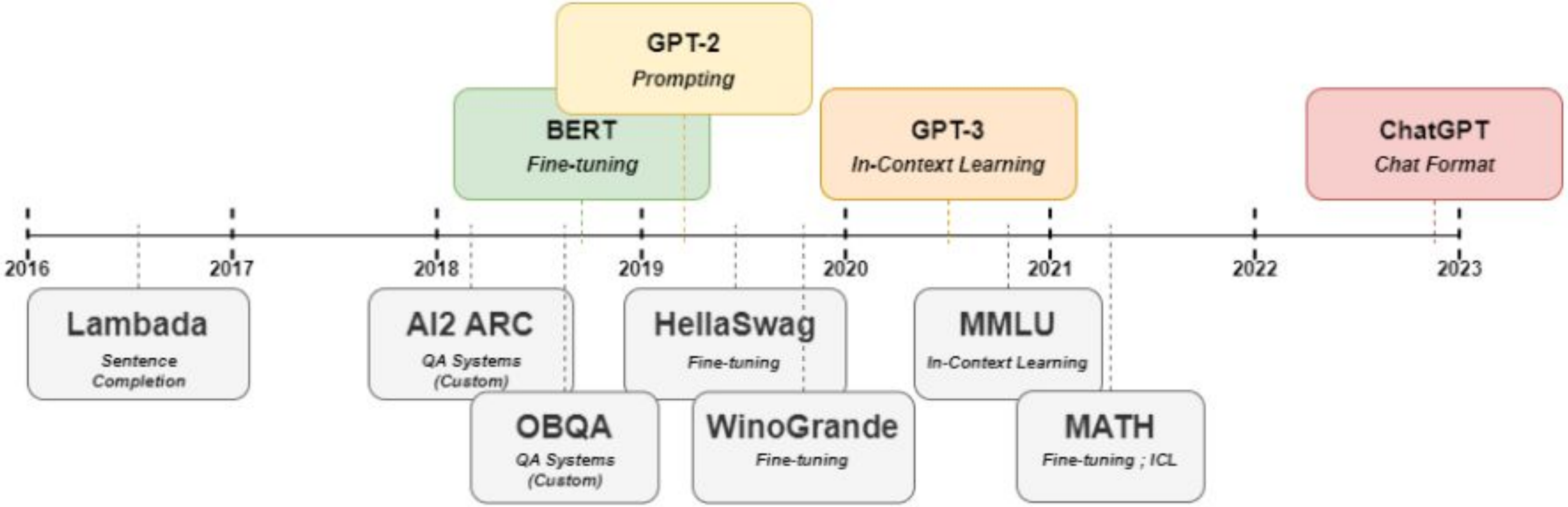
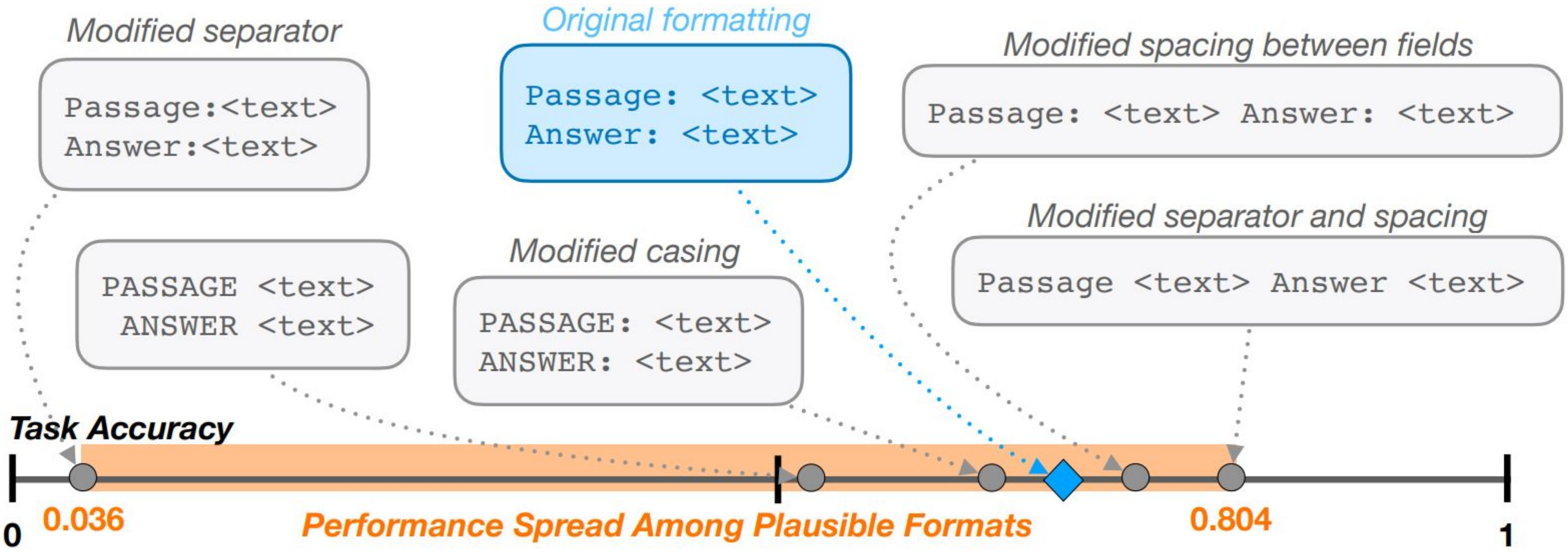


Figure 1: A timeline showing the relative release dates of a selection of notable benchmarks used to evaluate LMs, as compared to the release dates of BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), and ChatGPT, used as approximate stand-ins for shifts in how the community uses and therefore evaluates LMs. Common practice

https://arxiv.org/abs/2405.14782

Prompt specific issues

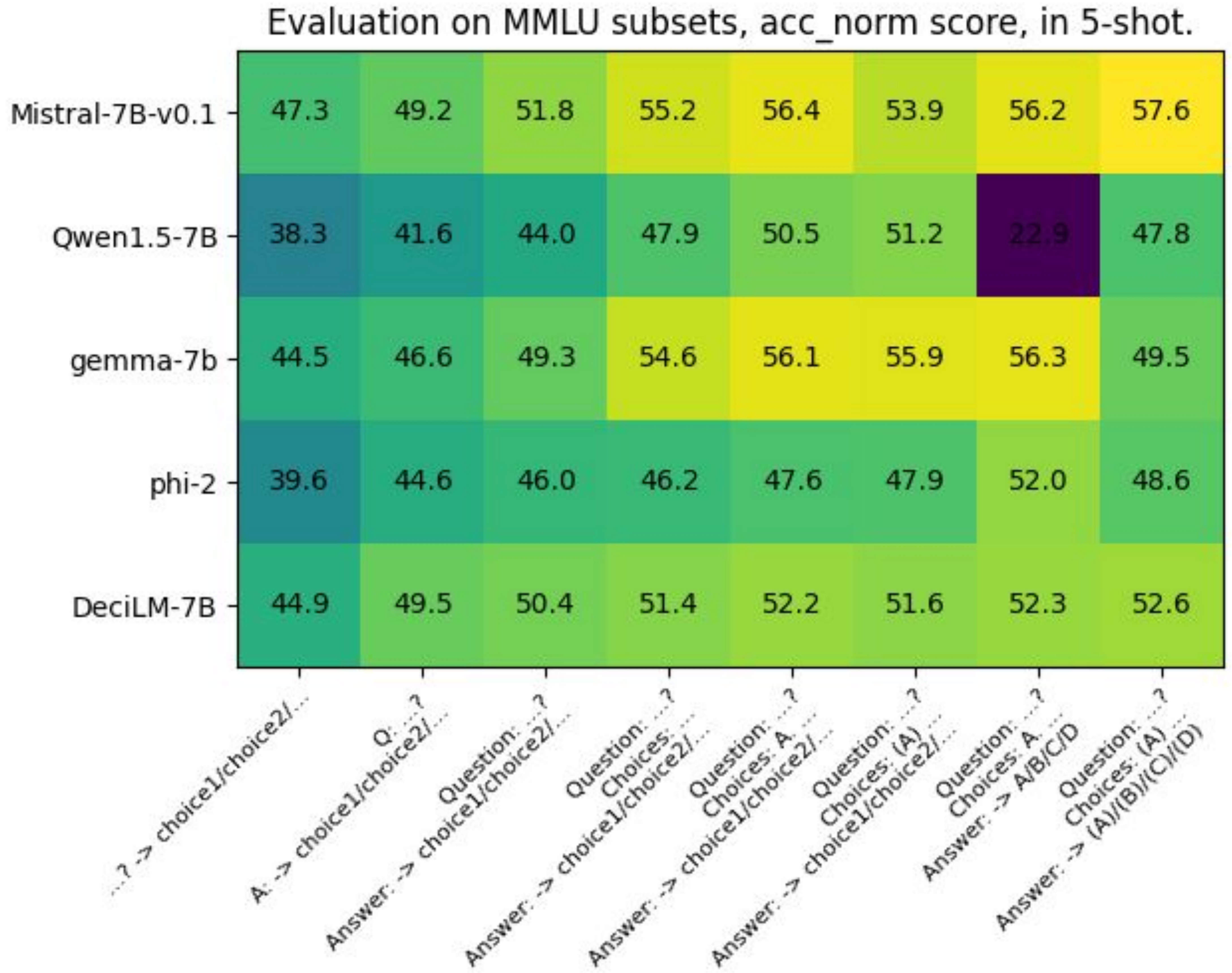
Sensitivity to prompt format

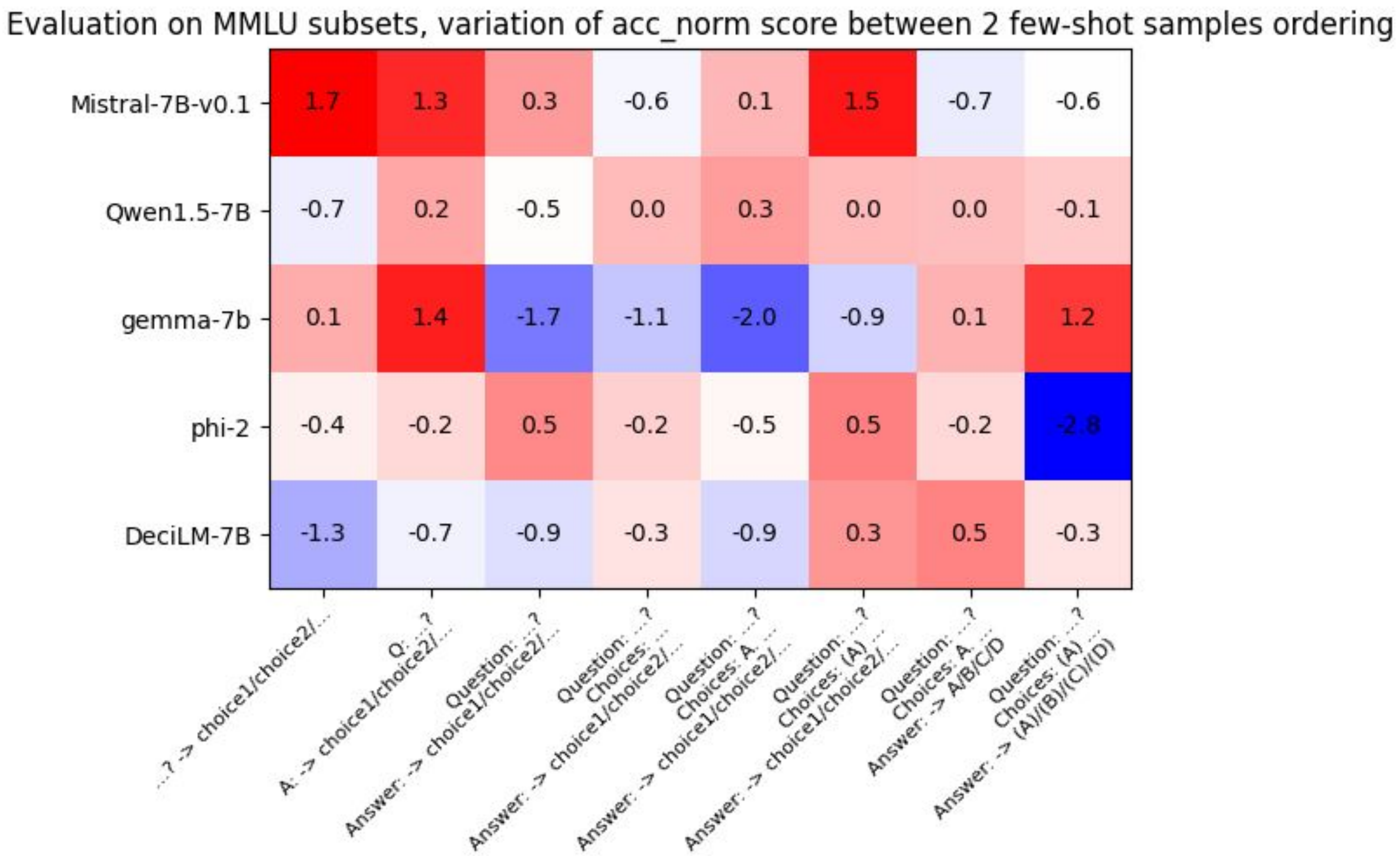


https://arxiv.org/abs/2310.11324

Prompt specific issues

Sensitivity to the prompt format or few shot ordering

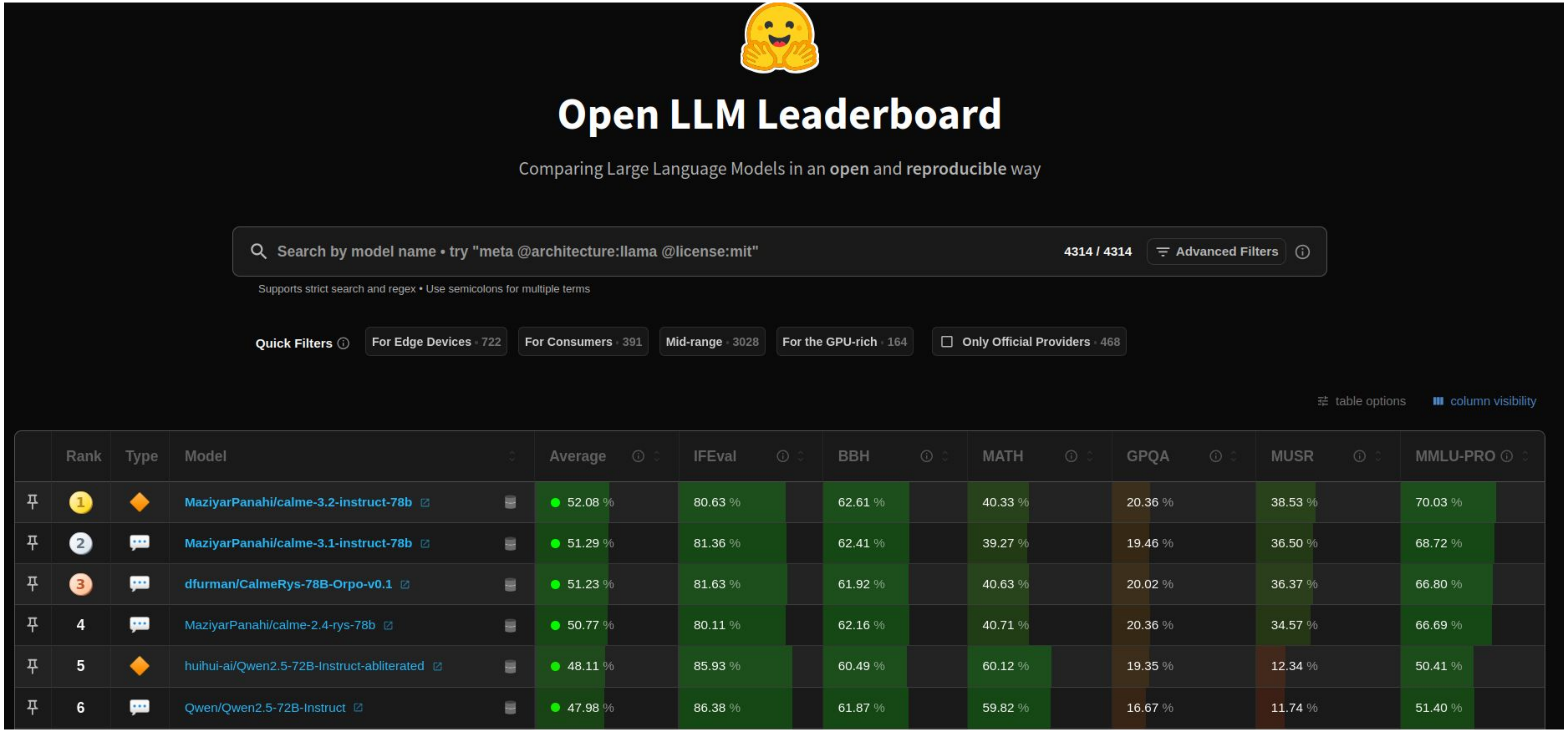




https://huggingface.co/blog/evaluation-structured-outputs

Evaluation in practice
Comparing models in the open: leaderboards

Open LLM Leaderboard: 13K models over 2 years



Open LLM Leaderboard

Comparing Large Language Models in an **open** and **reproducible** way

Search by model name • try "meta @architecture:llama @license:mit" 4314 / 4314 Advanced Filters

Supports strict search and regex • Use semicolons for multiple terms

Quick Filters: For Edge Devices · 722 | For Consumers · 391 | Mid-range · 3028 | For the GPU-rich · 164 | Only Official Providers · 468

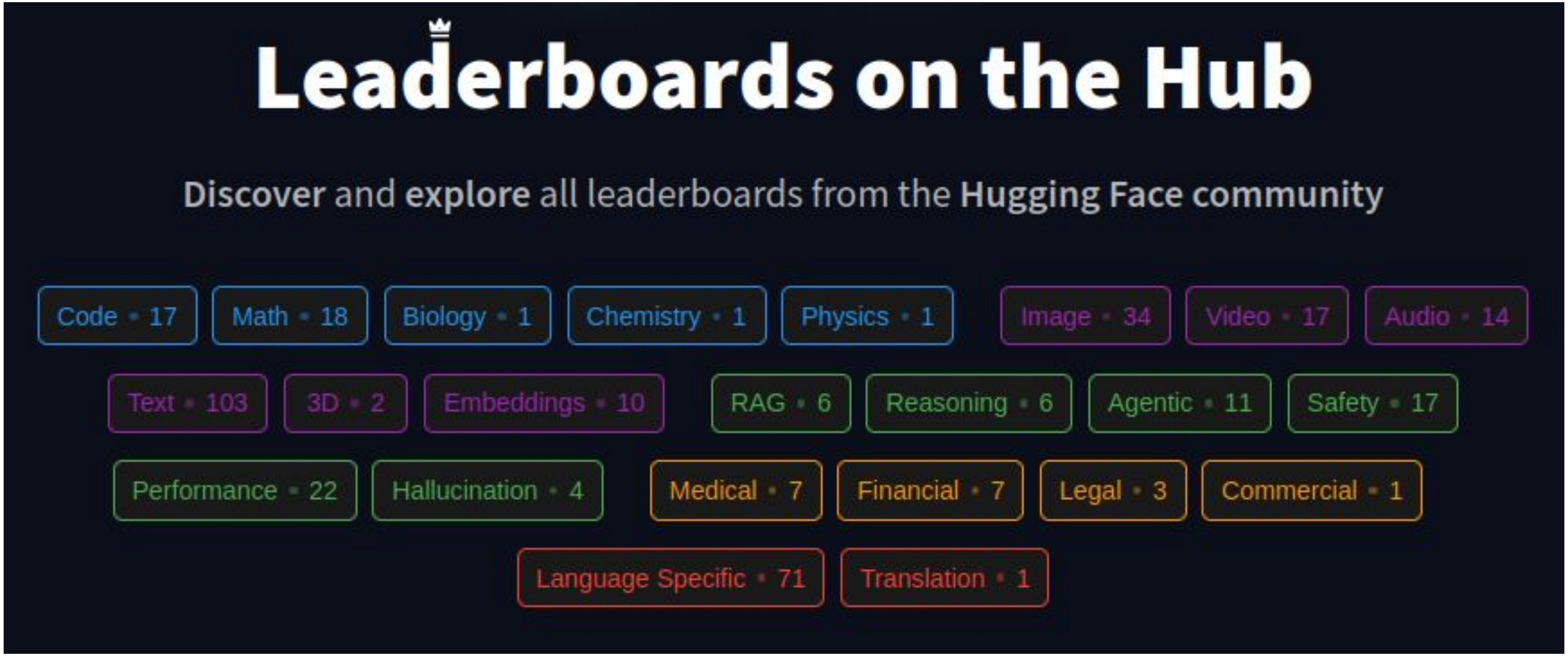
table options | column visibility

Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU-PRO
1		MaziyarPanahi/calme-3.2-instruct-78b	52.08 %	80.63 %	62.61 %	40.33 %	20.36 %	38.53 %	70.03 %
2		MaziyarPanahi/calme-3.1-instruct-78b	51.29 %	81.36 %	62.41 %	39.27 %	19.46 %	36.50 %	68.72 %
3		dfurman/CalmeRys-78B-Orpo-v0.1	51.23 %	81.63 %	61.92 %	40.63 %	20.02 %	36.37 %	66.80 %
4		MaziyarPanahi/calme-2.4-rys-78b	50.77 %	80.11 %	62.16 %	40.71 %	20.36 %	34.57 %	66.69 %
5		huihui-ai/Qwen2.5-72B-Instruct-abliterated	48.11 %	85.93 %	60.49 %	60.12 %	19.35 %	12.34 %	50.41 %
6		Qwen/Qwen2.5-72B-Instruct	47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %

https://huggingface.co/open-llm-leaderboard/

Clémentine Fourrier

Leaderboards on the Hub: 200 community-led benchmarks



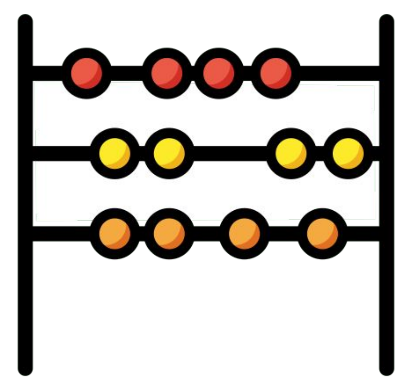
https://huggingface.co/spaces/OpenEvals/find-a-leaderboard

Evaluation in practice

Knowing where we are going

Evaluations to follow this year

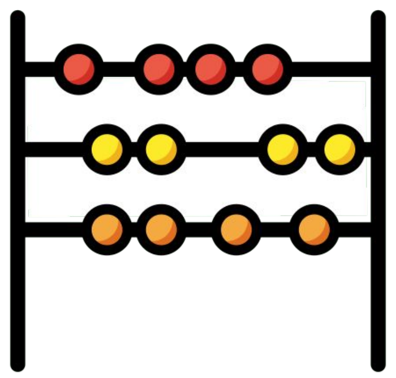
AIME/Frontier Math



AIME - American Invitational Mathematics Examination
- High school level olympiad math problem solving
- Fully public, annually updated
- Max scores: ~30 to 40% (on 2025 and 2024 editions)

FrontierMath
- Expert level math problems, written by hand
  - novel + unpublished + verifiable/guessproof + verified
- Fully private, possible contamination of Open AI models
- Max scores: ~2% (25% for OpenAI o3)

FrontierMath example



Testing Artin's primitive root conjecture

Problem Solution

For a positive integer n, let $v_p(n)$ denote the largest integer v such that $p^v \mid n$. For a prime p and $a \not\equiv 0 \pmod{p}$, let $\operatorname{ord}_p(a)$ denote the smallest positive integer o such that $a^o \equiv 1 \pmod{p}$. For $x > 0$, let

$$\operatorname{ord}_{p,x}(a) = \prod_{\substack{q \le x \\ q \text{ prime}}} q^{v_q(\operatorname{ord}_p(a))} \prod_{\substack{q > x \\ q \text{ prime}}} q^{v_q(p-1)}.$$

Let S_x denote the set of primes p for which

$$\operatorname{ord}_{p,x}(2) > \operatorname{ord}_{p,x}(3),$$

and let d_x denote the density

$$d_x = \frac{|S_x|}{|\{p \le x : p \text{ is prime}\}|}$$

of S_x in the primes. Let

$$d_\infty = \lim_{x \to \infty} d_x.$$

Compute $\lfloor 10^6 d_\infty \rfloor$.

SWE-Bench Verified/SWE-Arena

SWE-Bench
- Issue-pull request pairs from github: models have to generate code which solves the post PR behavior
- Verified subset: manually annotated
- Max scores: ~50%

SWE-Arena
- “Battle” of code model across languages and tasks
- Includes a sandbox
- Associated leaderboard not out yet

SWE-Bench example

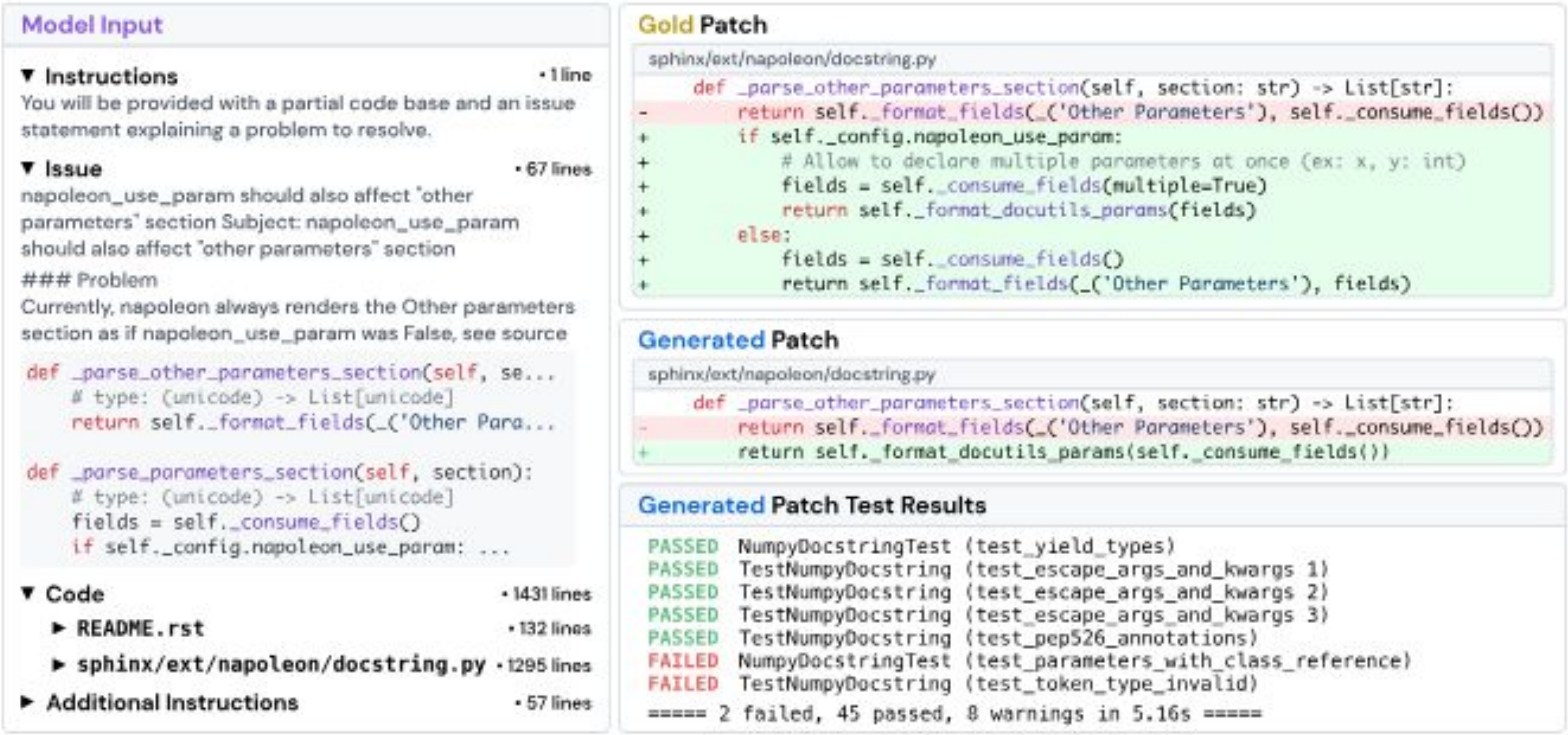


Figure 6: We show an example of an formatted task instance, a model prediction, and the testing framework logs. In the patches, red highlights are deletions. Green highlights are additions.

GPQA/HLE

Google Proof graduate Question Answers

- PhD level knowledge questions in chemistry, physics, biology
- Public
- Max scores: ~70%

Humanity's last exam

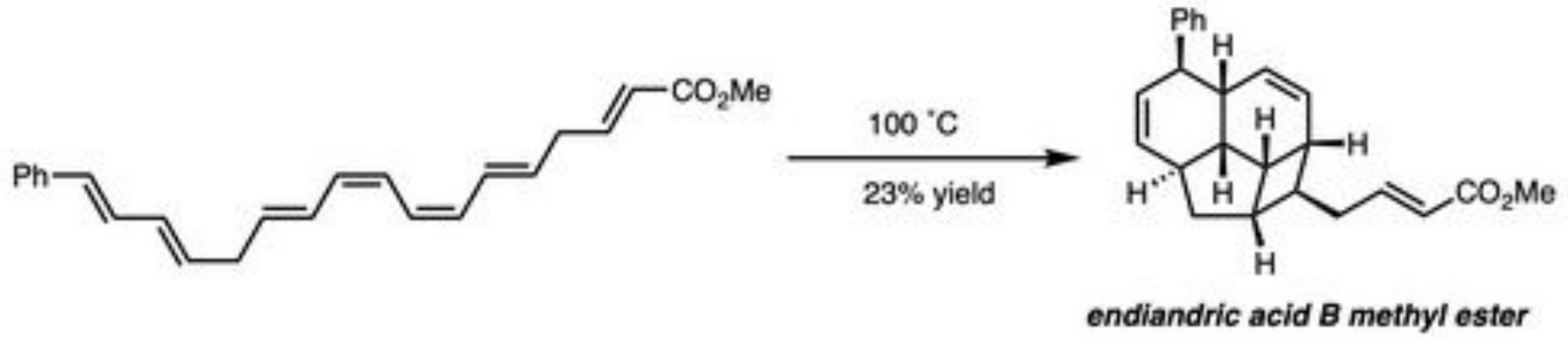
- Expert level knowledge questions across topics (sometimes require reasoning)
- Multimodal
- Max scores: ~10%

Humanity's last exam examples

Question:

Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYNᵓ BT ḤRY BR ʿTᵓ ḤBL

Question:



The reaction shown is a thermal pericyclic cascade that converts the starting heptaene into endiandric acid B methyl ester. The cascade involves three steps: two electrocyclizations followed by a cycloaddition. What types of electrocyclizations are involved in step 1 and step 2, and what type of cycloaddition is involved in step 3?

Provide your answer for the electrocyclizations in the form of [nπ]-con or [nπ]-dis (where n is the number of π electrons involved, and whether it is conrotatory or disrotatory), and your answer for the cycloaddition in the form of [m+n] (where m and n are the number of atoms on each component).

SciCode/DAB Step

SciCode

- Code generation problems to solve realistic scientific research problems, in Python
- Public
- Max scores: ~5% on the main problems

Data Agent Benchmark Step

- Data analysis problems on real life data requiring multistep problem solving
- Questions public, answers private
- Max scores: ~16% on the hard set, 73% on the easy set

SciCode example

Main Problem

Question: Generate an array of Chern numbers for the Haldane model on a hexagonal lattice by sweeping the following parameters: [MORE QUESTION TEXT]

Docstrings

```
def compute_chern_number_grid(delta, a, t1, t2, N):
    """
    Args:
    delta (float): The grid size in kx and ky axis.
    [MORE ARGUMENTS]

    Returns:
    results (ndarray): 2D array of shape(N, N), the Chern numbers.
    [MORE RETURN VALUES]
    """
```

Dependencies

```
import numpy as np
import cmath
from math import pi, sin, cos, sqrt
```

Subproblem 1

Background: Source: [CITATION]
$\{\mathbf{a}_i\}$ are the vectors from a B site to its three nearest-neighbor A sites, then we have [MORE BACKGROUND TEXT]

Question: Write a Haldane model Hamiltonian on a hexagonal lattice.

Docstrings

```
def calc_hamiltonian(kx, ky, a, t1, t2, phi, m):
    """
    Function to generate the Haldane Hamiltonian.

    Args:
    kx (float): The x component of the wavevector.
    [MORE ARGUMENTS]

    Returns:
    hamiltonian (ndarray): matrix of shape(2, 2).
    """
```

Subproblem 2

Background: Source: [CITATION]
Here we can discretize the two-dimensional Brillouin zone into grids with step [MORE BACKGROUND TEXT]

Question: Calculate the Chern number using the Haldane Hamiltonian.

Docstrings

```
def compute_chern_number(delta, a, t1, t2, phi, m):
    """
    Function to compute the Chern number.

    Args:
    delta (float): The grid size in kx and ky axis.
    [MORE ARGUMENTS]

    Returns:
    chern_number (float): The Chern number.
    """
```

Subproblem 3

Question: Here we can discretize the two-dimensional Brillouin zone into grids with step [MORE QUESTION TEXT]

Docstrings

```
def compute_chern_number_grid(delta, a, t1, t2, N):
    """
    Function to calculate the Chern numbers.

    Args:
    delta (float): The grid size in kx and ky axis for discretizing the
Brillouin zone.
    [MORE ARGUME]

    Returns:
    results (ndarray): 2D array of shape(N, N), The Chern numbers.
    [MORE RETURN VALUES]
    """
```

GAIA - General AI Assistant benchmark

- Questions requiring a combination of tool use, multistep reasoning, and multimodality to solve
- Questions public, answers private
- Max scores: ~40% for the level 3 questions

GAIA examples

Level 1

Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?
Ground truth: 90

Level 2

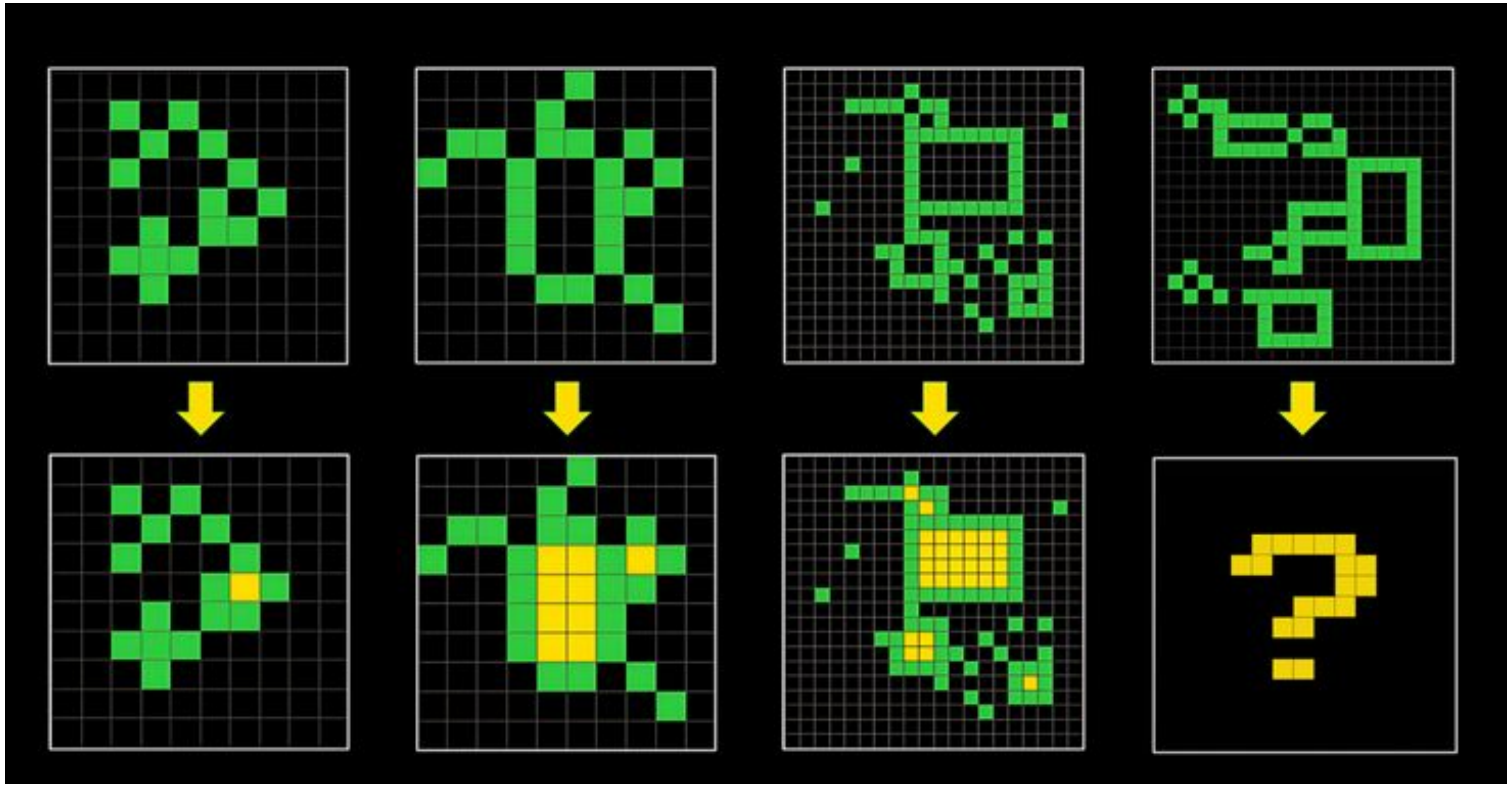
Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.
Ground truth: +4.6

Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes.
Ground truth: White; 5876

ARC-AGI

- Puzzle like grid completion challenges requiring pattern matching/reasoning
- Private
- Max scores: ~53%



Last thoughts

Open thoughts & questions

- Evals are only interesting if they are hard - Saturation
- Rankings only hold as long as everyone plays fair - Contamination
- Comparing to humans make little sense - Baselines
- Making sure an eval is a good proxy for a capability is hard
- Are we looking in the correct direction?

Trends to follow in 2025

- Synthetic evaluations
  - Custom use cases
- Shift in evaluations topics
  - Agentic
- Performance evaluation focus
  - Inference cost
  - On device models
  - Environmental footprint

Questions

Clémentine Fourrier

OpenEvals team at