Evaluation

What the numbers measure, and when they lie

Required

Zheng et al., Judging LLM-as-a-Judge · arXiv:2306.05685

Miller, Adding Error Bars to Evals · arXiv:2411.00640

Optional

Zhang et al., GSM1k · arXiv:2405.00332 Bhagia et al., Model Ladders · arXiv:2412.04403

Chiang et al., Chatbot Arena · arXiv:2403.04132 Magnusson et al., Paloma · arXiv:2312.10523

Christoffer Heckman

Department of Computer Science, University of Colorado Boulder

September 3, 2026

0 FIFTEEN MINUTES

Quiz 1

Quick overview

Your number. Assignment 1 trains to a val loss near 1.46. Is that model good?

Teacher forcing, and what loss means off its home domain

Loss to capability. Why bigger models score better on benchmarks at all.

The ladder your Part 3 fit belongs to

The benchmarks. What MMLU, GSM8K, HumanEval, SWE-bench and Arena each measure.

And the scoring rules underneath them

When numbers lie. Contamination, fragile rankings, biased judges, thin samples.

Then the harness that runs all of it: Assignment 2

1 YOUR NUMBER

Is 1.46 Good?

One Number

1.462

best val loss, base run, nats per character · Assignment 1 codebase

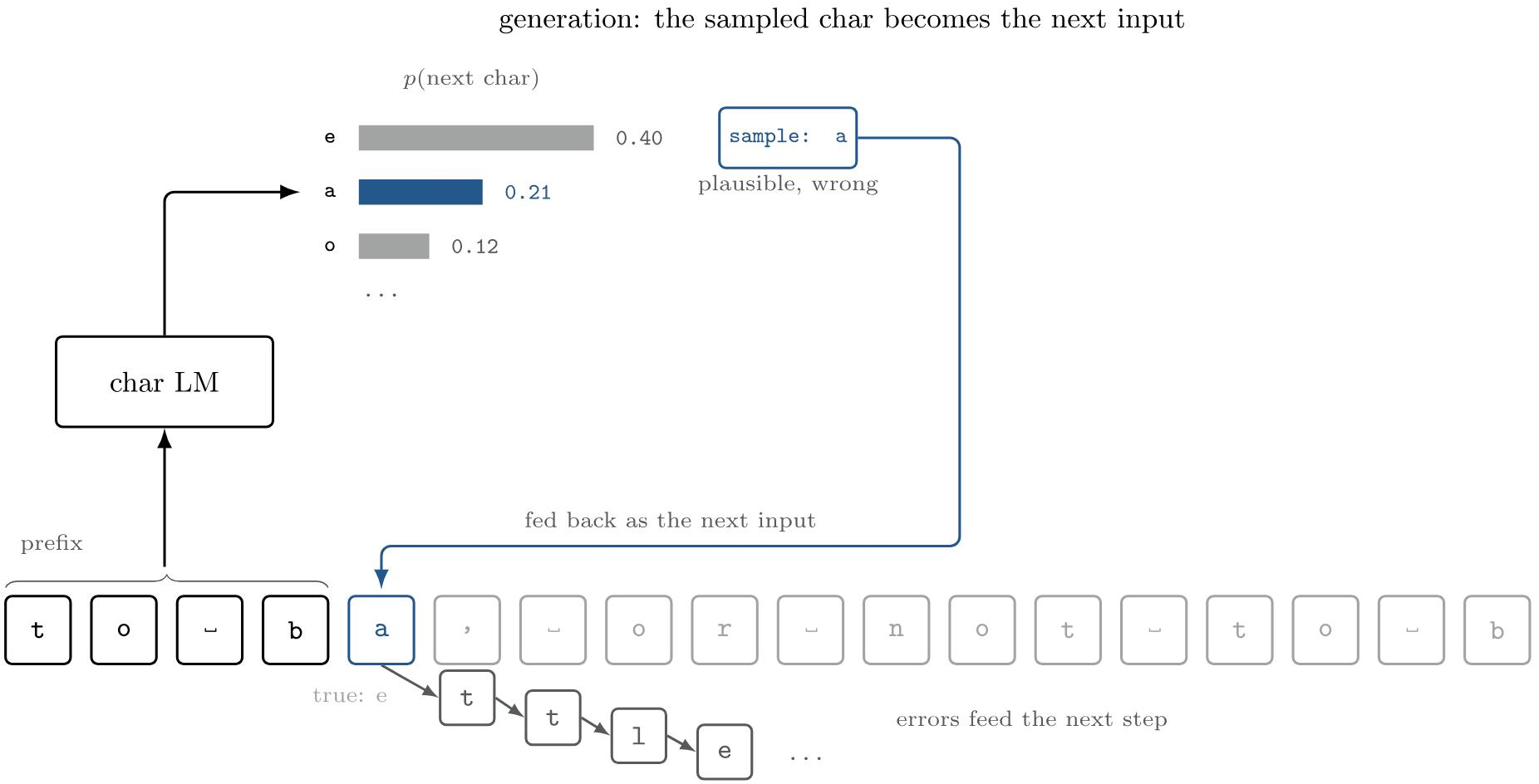
```
DUKE OF YORK:
What envious throne as we are,
Or was the body of births, the baby of sensess,
And hardly said you of shame, will not speak to grief.

DUCHESS OF YORK:
Stay and we there at Richard specially.

DUKE OF YORK:
Ay, and much in his book she not blood?
```

sampled from the same checkpoint, prompt "DUKE OF", temperature 0.8 (excerpt)

Teacher Forcing



Four Exams

Shakespeare, held out

```
GREMIO:

Good morrow, neighbour Baptista.
```

the val split from Assignment 1: same distribution, never trained on

TinyStories

```
Spot saw the shiny car and said, "Wow, Kitty, your

car is so bright and clean!"
```

synthetic children's stories in simple modern English; returns in Assignment 2

Wikipedia

```
World War II, or the Second World War, was a

global conflict between two coalitions.
```

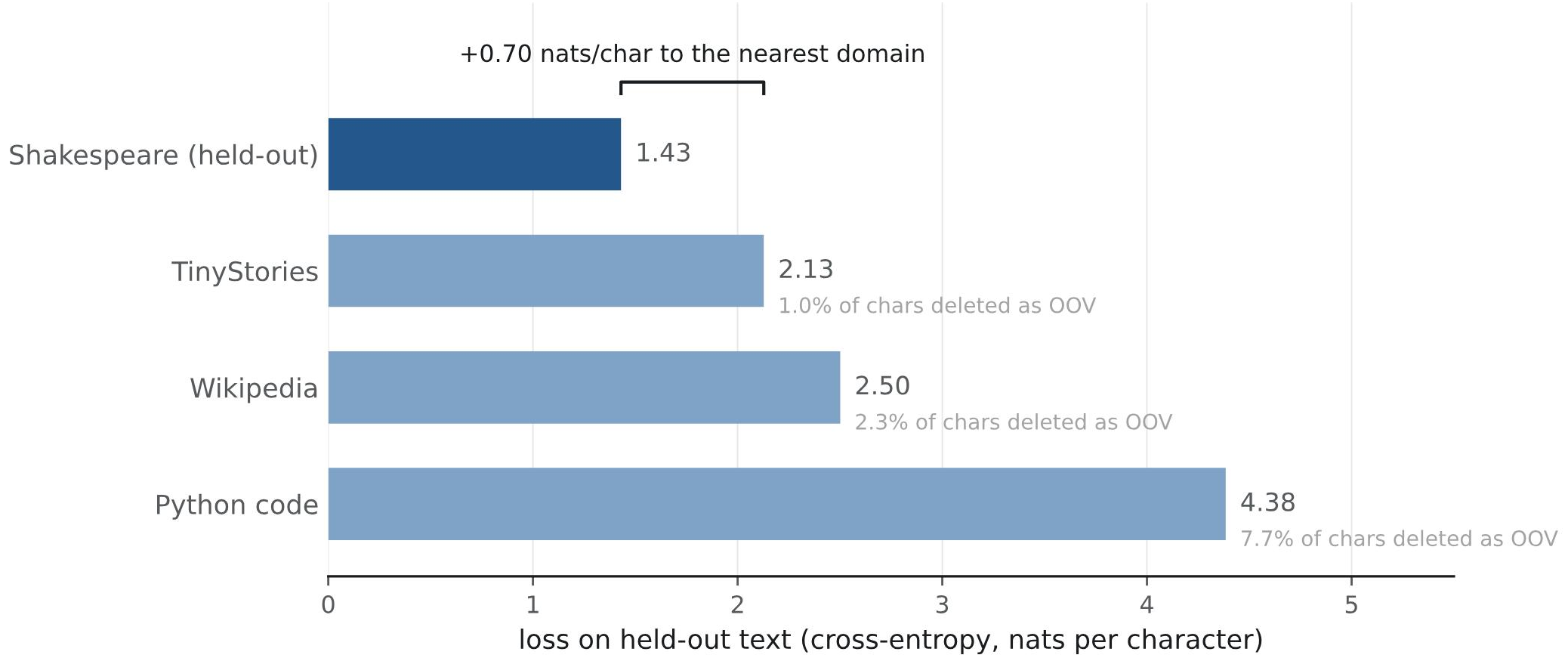
ten full articles of encyclopedic prose, mixed topics

Python source

```
def parse_args(self, args=None, namespace=None):
```

standard library code: argparse, ast, inspect and friends

Changing the Exam



cross-entropy: the minus log p loss from Lecture 1, averaged per character

out/base: your Assignment 1 base checkpoint, 10.6M parameters

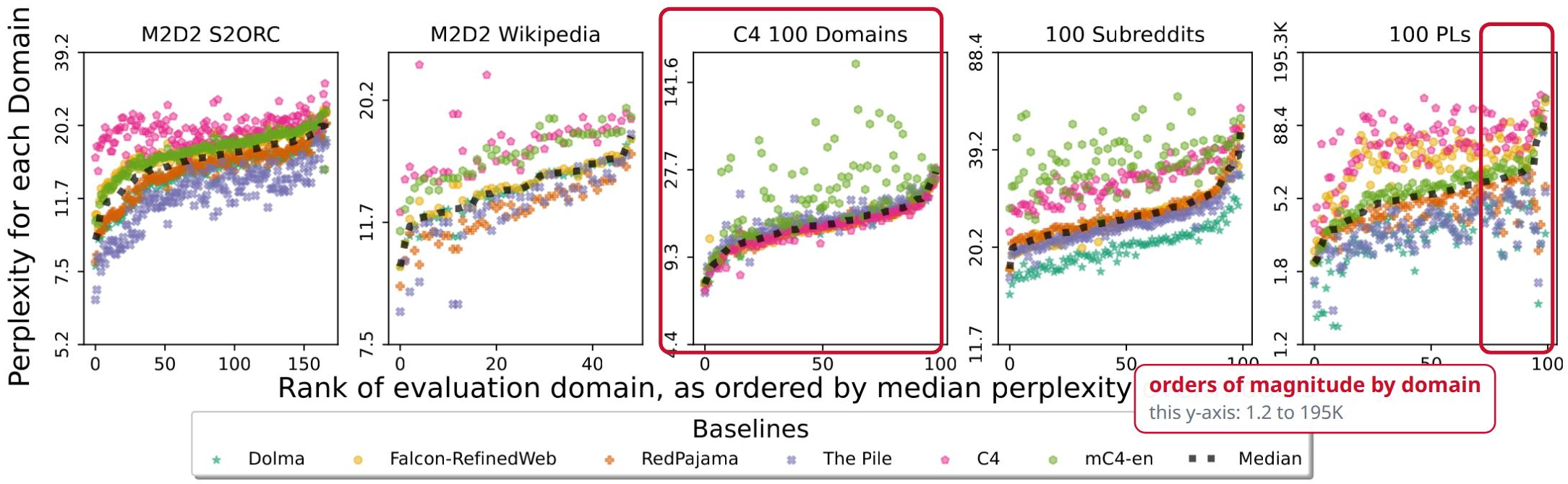
teacher-forced: scored with the true prefix, previous slide

Assignment 1 codebase, base checkpoint, block size 256. Every corpus is held-out evaluation text the model never trained on. Foreign corpora normalized to the 65-character vocabulary; unrepresentable characters deleted and counted.

Changing the Exam

Assignment 1 codebase, all five ladder rungs, same protocol as the previous slide.

A Vector, Not a Scalar



orders of magnitude by domain
this y-axis: 1.2 to 195K

same six models, one domain axis

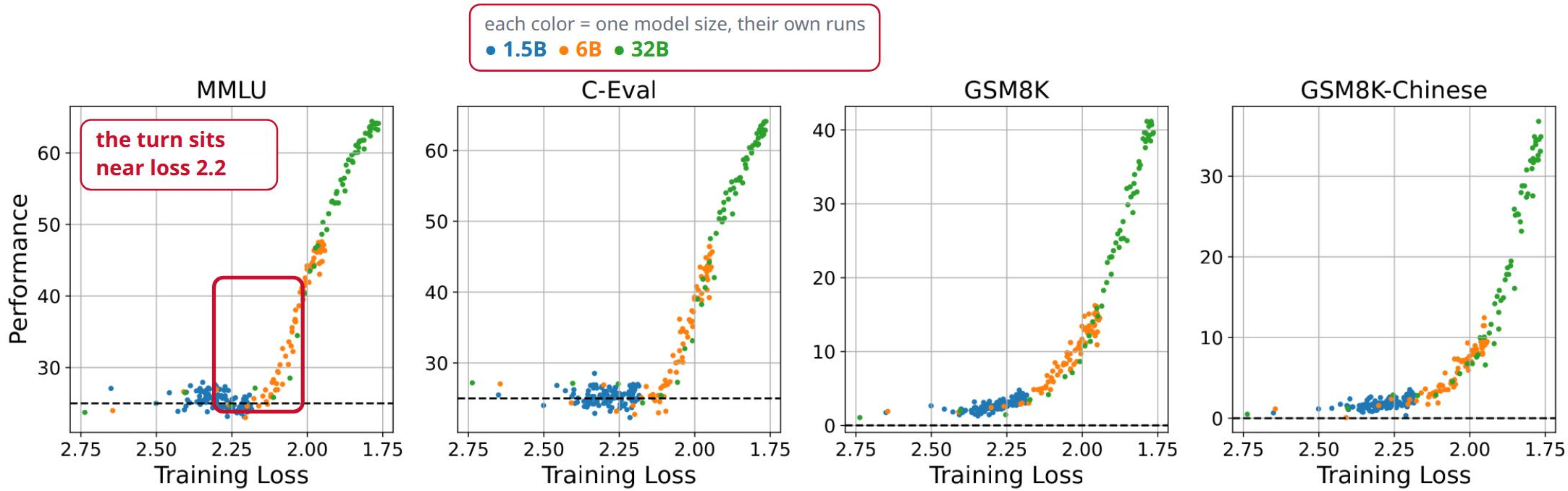
perplexity = e^{loss}: the number of choices the model is effectively guessing between. Our base run's 1.43 nats is perplexity 4.2.

Magnusson et al., "Paloma: A Benchmark for Evaluating Language Model Fit," 2023, arXiv:2312.10523, Figure 3.

2 LOSS TO CAPABILITY

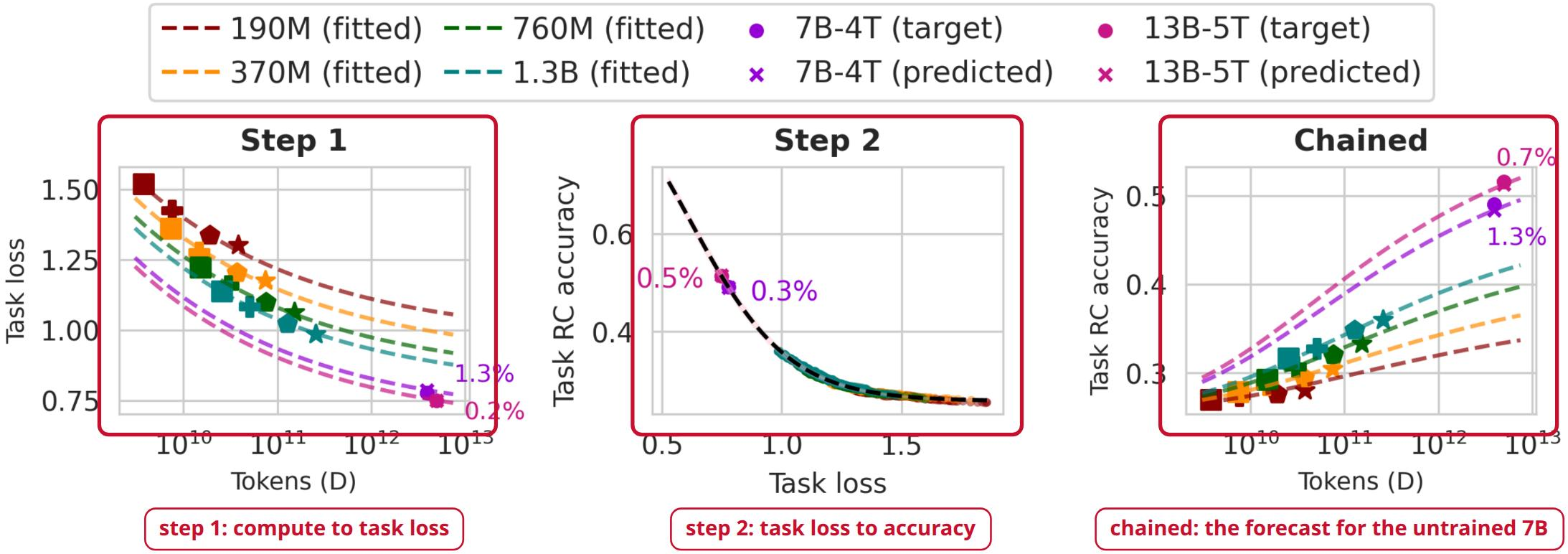
What Scale Buys

From Loss to Accuracy



Du et al., "Understanding Emergent Abilities of Language Models from the Loss Perspective," 2024, arXiv:2403.15796, Figure 1 (bottom row).

The Ladder, Again

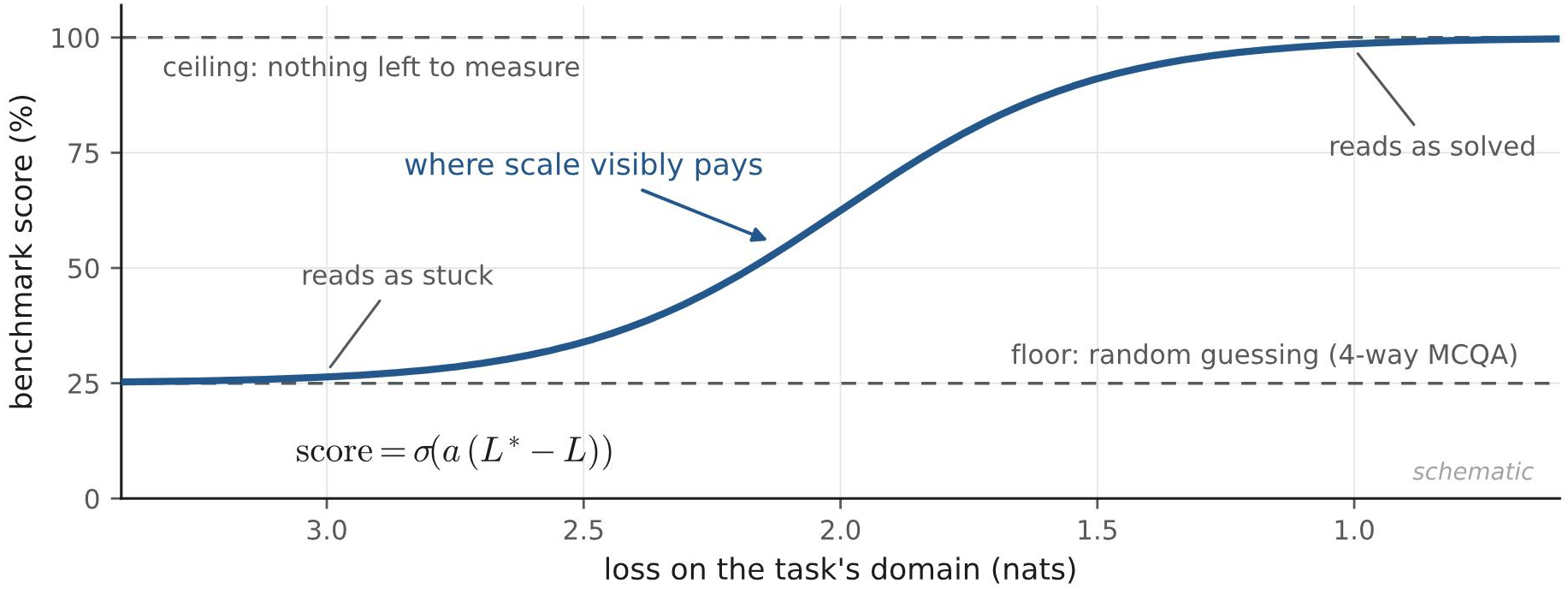


In the legend: **predicted** is the ladder fit's forecast, made before the big model existed. **target** is what the trained model then measured.

Bhagia et al., "Establishing Task Scaling Laws via Compute-Efficient Model Ladders," 2024, arXiv:2412.04403, Figure 1.

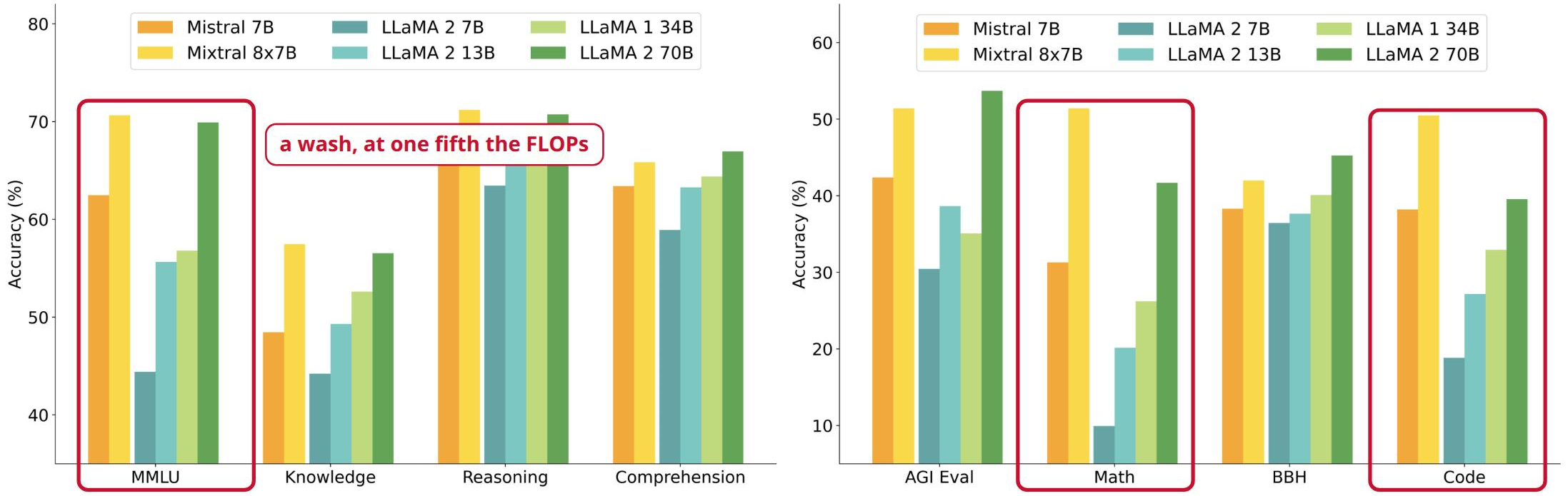
Why Bigger Helps

$$C \xrightarrow{\text{power law (L1)}} L_{\text{domain}} \xrightarrow{\text{link}} \text{score}$$



Scale helps on an eval as far as loss transfers to it, and no further.

Different Wins



Jiang et al., "Mixtral of Experts," 2024, arXiv:2401.04088, Figure 2.

3 THE BENCHMARKS

Scoring Rules

The Landscape

columns: what is measured rows: how it is scored	Knowledge	Math	Code	Agentic software	Open-ended dialogue
Choice log-likelihood (MCQA)	MMLU MMLU-Pro GPQA				
Exact match after CoT		GSM8K MATH AIME			
Execution pass@k			HumanEval MBPP		
Resolved rate				SWE-bench	
Pairwise preference					MT-Bench Chatbot Arena

Multiple Choice

$$\hat{a} \;=\; \arg\max_i \;\log p_\theta(\text{choice}_i \mid \text{question})$$

Score the letter or score the full answer text, with or without length normalization. The same model on the same questions reports different accuracies.

A 4-way choice has a floor of 25. And a model can read the choices without reading the question.

pass@k

$$\text{pass@}k \;=\; \mathbb{E}\left[\,1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\,\right]$$

n samples drawn, c passed the tests. The naive $1-(1-\frac{c}{n})^k$ reads high.

$$n = 200,\ c = 13: \qquad \text{pass@}1 = 6.5\%, \qquad \text{pass@}10 = 49.8\%$$

Execution is the answer key: the tests define correctness, so the tests bound what the score means.

Chen et al., "Evaluating Large Language Models Trained on Code," 2021, arXiv:2107.03374, Section 2.1 (the unbiased pass@k estimator).

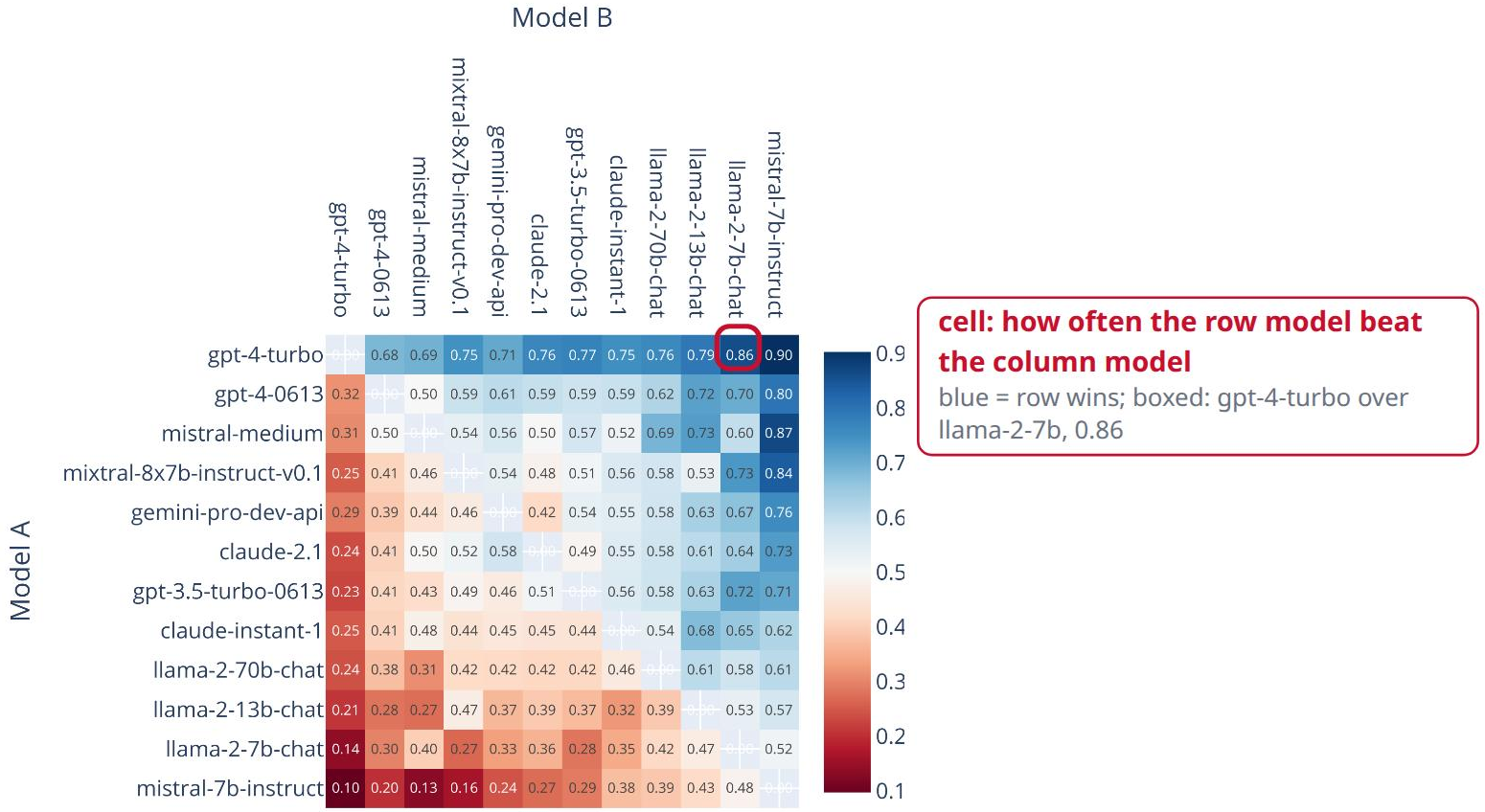
4 PREFERENCES AND JUDGES

No Answer Key

Preferences

$$P(i \succ j) = \frac{1}{1 + e^{-(\theta_i - \theta_j)}}$$

a human sees two anonymous answers: "which do you prefer?"



rows sorted by fitted rating, best first

so the matrix shades blue toward the top-right

cell: how often the row model beat the column model

blue = row wins; boxed: gpt-4-turbo over llama-2-7b, 0.86

The rating measures what voters reward, and style is part of what they reward.

Chiang et al., "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," 2024, arXiv:2403.04132, Figure 2 (left panel).

The Judge

```
You are grading two responses to the same instruction.
Criteria: helpfulness, correctness, depth. Compare A and B.
Verdict: "A", "B", or "tie".
```

Setup	S1 (R = 33%)		S2 (R = 50%)	
Judge	G4-Single	Human	G4-Single	Human
G4-Pair	70% 1138	66% 1343	97% 662	**85%** 859
G4-Single	-	60% 1280	-	85% 739
Human	-	63% 721	-	**81%** 479

(a) First Turn

the judge, 85 on non-tie votes

humans with each other, 81

Reading the table: **G4-Pair / G4-Single** is GPT-4 judging pairwise or grading one answer. **S1** counts tie votes, **S2** drops them. **R** is the agreement two random voters would reach: 33% with ties, 50% without.

Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," NeurIPS 2023, arXiv:2306.05685, Table 5(a), first turn; S2 counts non-tie votes.

Failure Modes

The judge’s

Position bias. Swap A and B, and the verdict changes.

Verbosity bias. Longer scores higher, content held equal.

Self-preference. The judge favors text in its own voice.

The benchmark's

Contamination. The test set leaked into the training data.

Format fragility. Reformat the questions, and the scores move.

Saturation. Scores crowd the ceiling; the gaps left are noise.

Judging the Judge

Claude v1, 23.8: below random; the order IS the verdict

Judge	Prompt	Consistency	Biased toward first	Biased toward second	Error
Claude-v1	default	23.8%	**75.0%**	0.0%	1.2%
	rename	56.2%	11.2%	**28.7%**	**3.8%**
GPT-3.5	default	46.2%	**50.0%**	1.2%	2.5%
	rename	51.2%	38.8%	6.2%	**3.8%**
GPT-4	default	**65.0%**	30.0%	5.0%	0.0%
	rename	**66.2%**	28.7%	5.0%	0.0%

GPT-4: same winner after the swap, 65% of pairs
judged twice, A and B trading places; random verdicts agree 33%

Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," NeurIPS 2023, arXiv:2306.05685, Table 2.

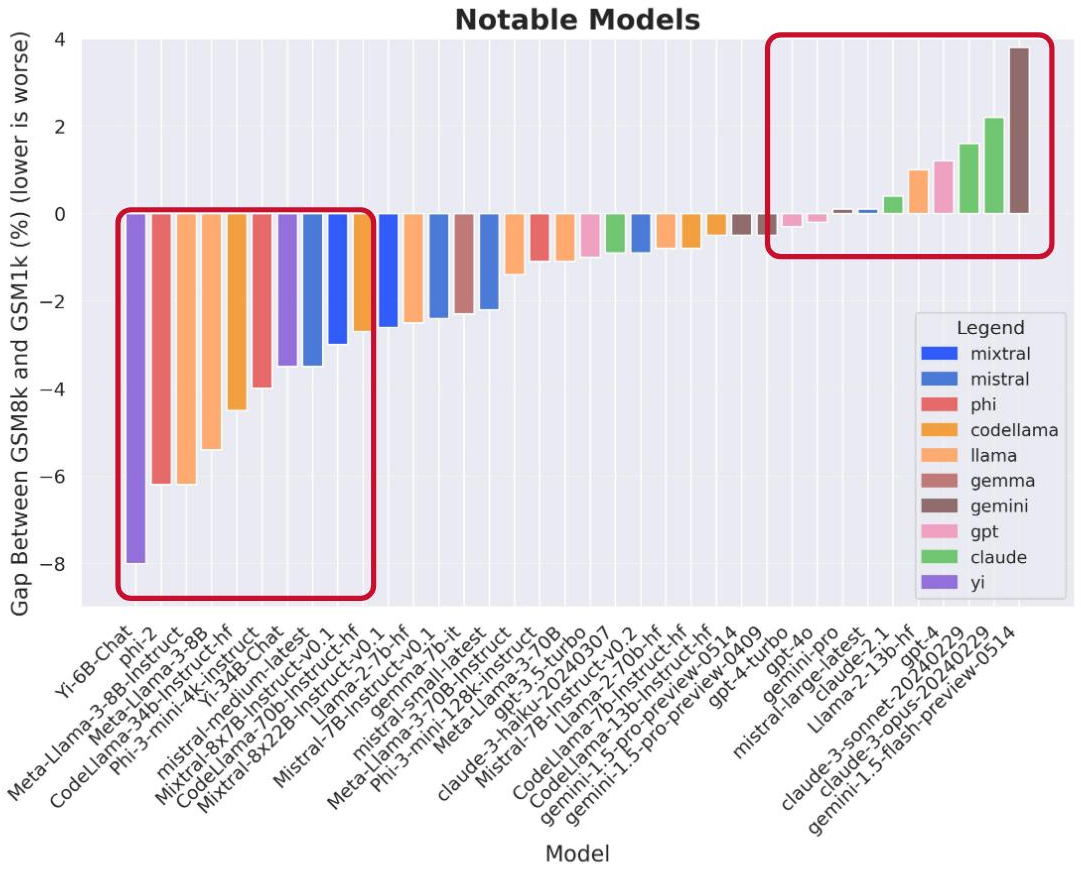
5 WHEN NUMBERS LIE

Memory, Format, Sample Size

Contamination

the setup: write 1,250 NEW problems at GSM8k difficulty; run every model on both tests; each bar = old score minus new score

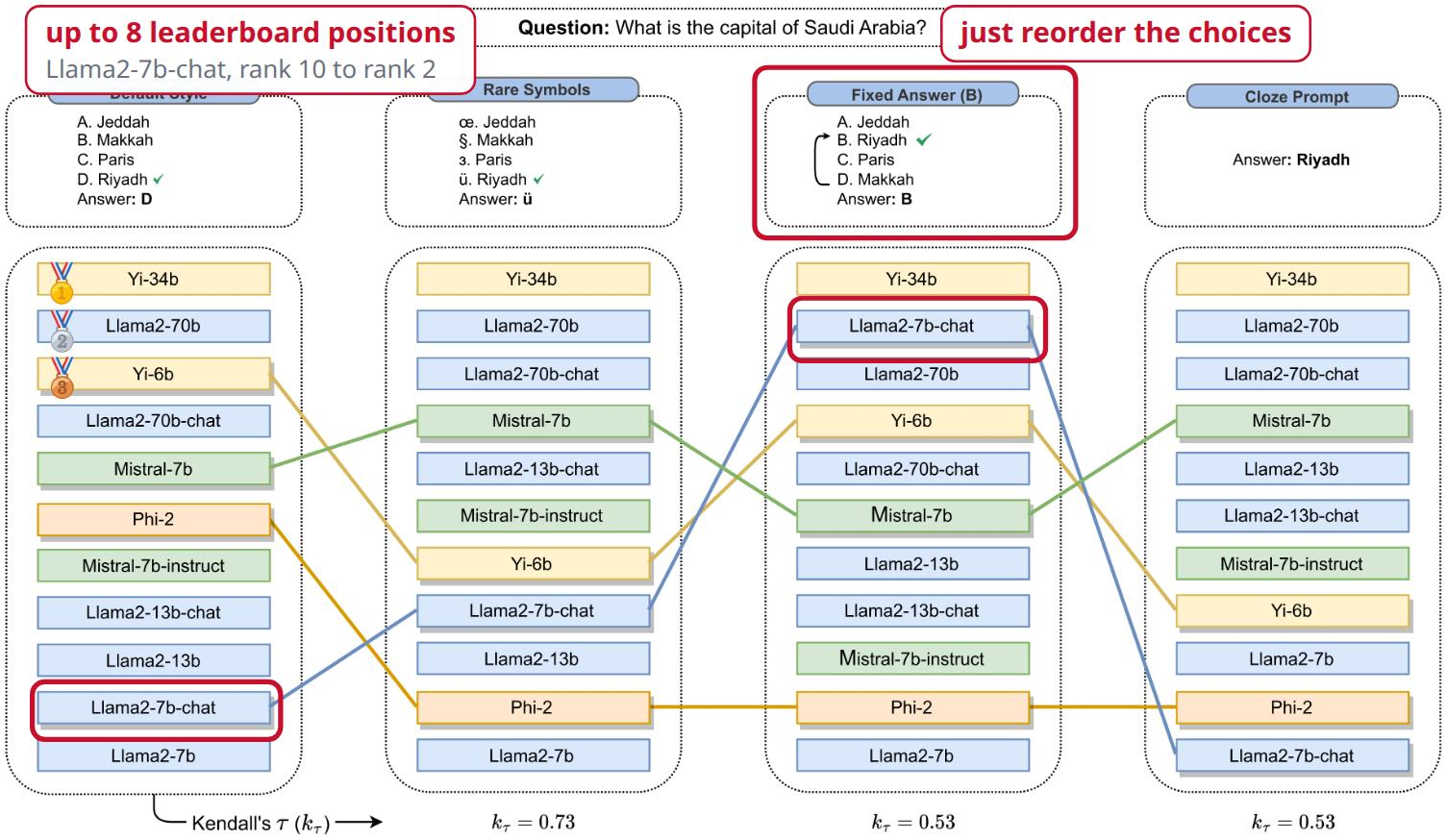
Phi and Mistral families: up to 8 points lower on the fresh problems
that gap is memorized test data, not math



frontier models: bar near zero, the skill is real

Zhang et al., "A Careful Examination of Language Model Performance on Grade School Arithmetic," 2024, arXiv:2405.00332, Figure 1.

Fragile Rankings



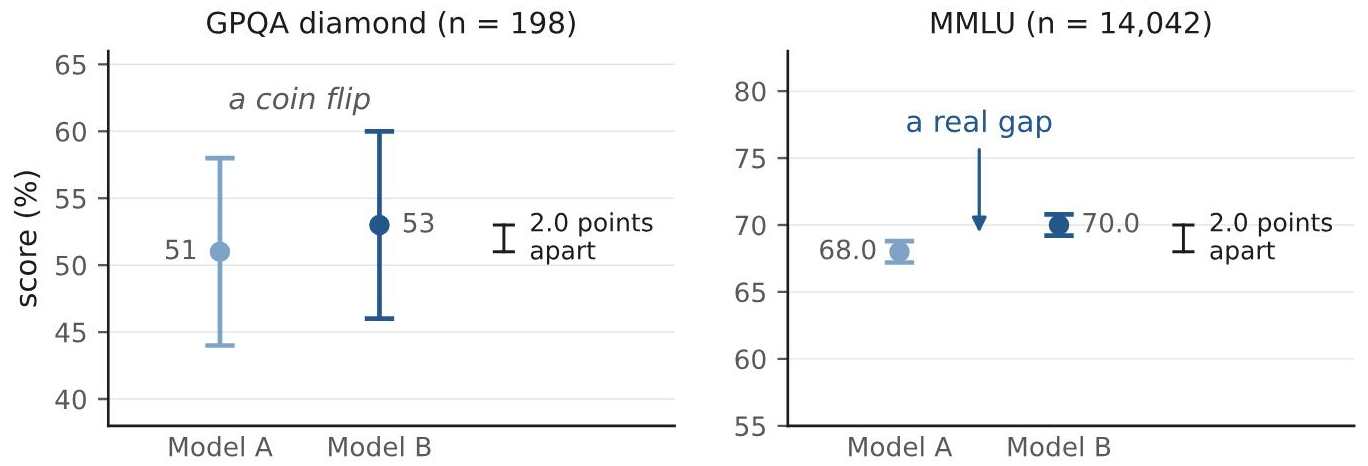
Alzahrani et al., "When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards," 2024, arXiv:2402.01781, Figure 1.

Error Bars

A benchmark score is an average over n questions, so it carries a standard error: $\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}$

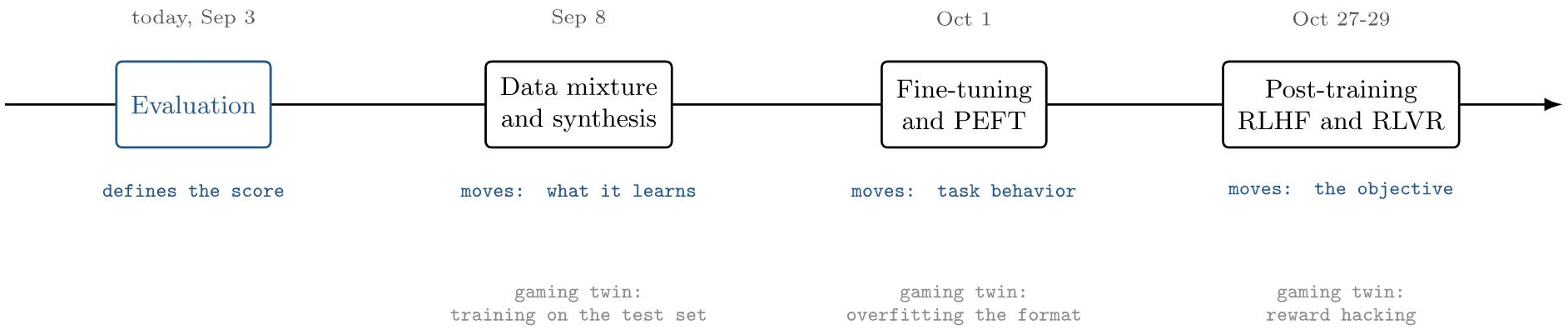
MMLU, $14{,}042$ questions: $\pm 1.96\sqrt{\frac{0.7 \cdot 0.3}{14{,}042}} = 0.8$ percentage points

GPQA diamond, 198 questions: $\pm 1.96\sqrt{\frac{0.5 \cdot 0.5}{198}} = 7.0$ percentage points



A two-point lead on a small benchmark is a coin flip. Paired, per-question comparison tightens it.

Making the Number Move



6 THE HARNESS

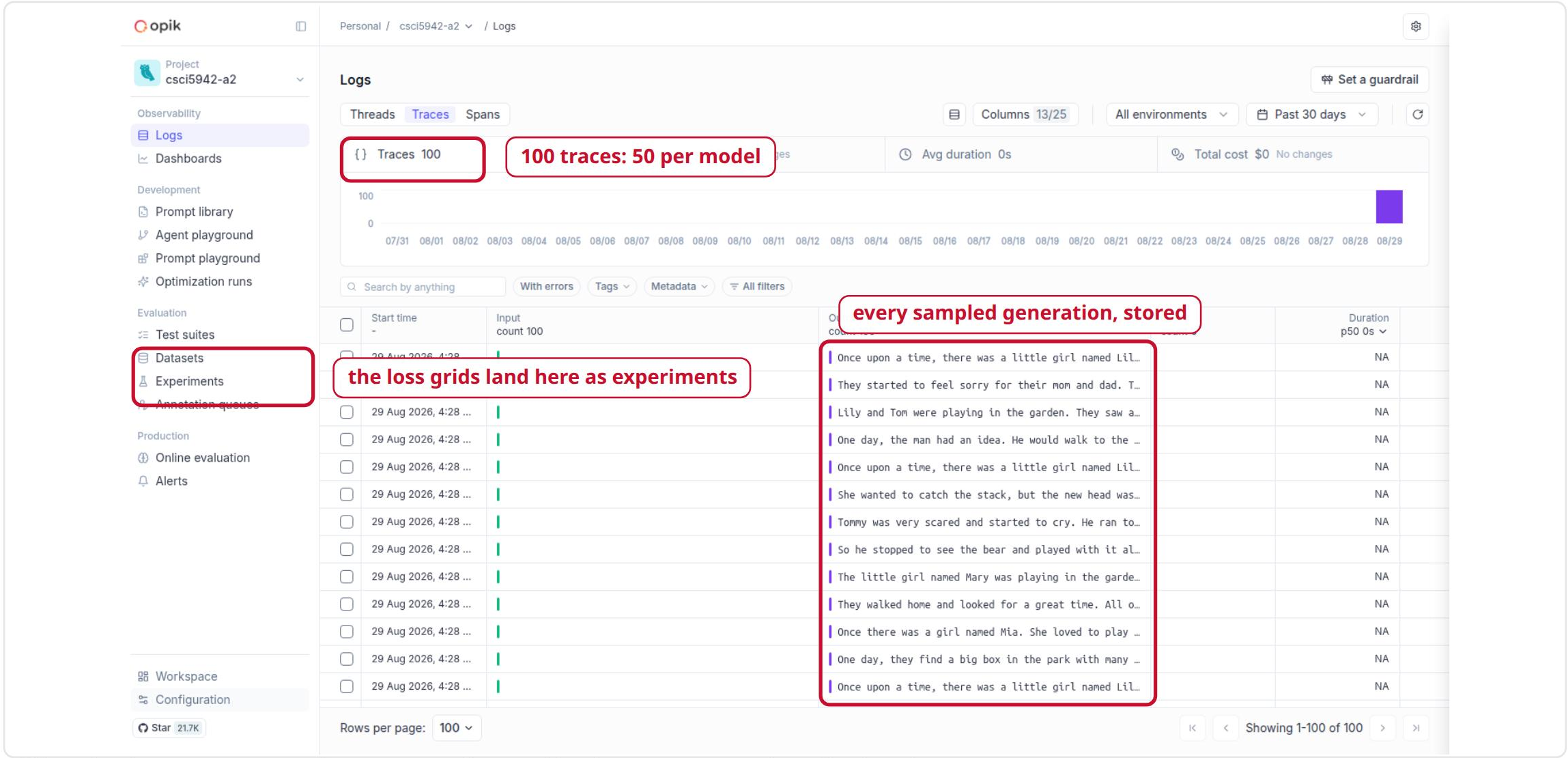
Evaluation as Infrastructure

Harnesses

```
task: mmlu_abstract_algebra
  dataset_path: cais/mmlu
  num_fewshot: 5
  doc_to_choice: ["A", "B", "C", "D"]
  metric: acc
```

The prompt, the shots, the scoring and the aggregation are pinned in config, or the number is not comparable. The same LLaMA 65B weights scored 63.7 on MMLU in one harness and 48.8 in another.

Opik



Self-hosted Opik, project csci5942-a2: the Assignment 1 base model and its TinyStories twin, logged by the A2 harness.

Assignment 2

Train the base config, unchanged, on a second corpus: TinyStories.

Same 82M characters of budget, two minutes on your A1 GPU

Evaluate both models on four domains: the loss grid from this lecture, yours.

Shakespeare, TinyStories, Wikipedia, Python, with the OOV accounting

Judge both models’ generations in Opik, with the order-swap control.

Gemini as judge, per-corpus rubric, error bars required

Deliverable: which model is better? Defend the answer per exam.

The correct answer names the exam it is true on

Reading

Required

Zheng et al., Judging LLM-as-a-Judge, arXiv:2306.05685. The judge you will run in A2, biases included.

Miller, Adding Error Bars to Evals, arXiv:2411.00640. The interval A2 asks for on every number you report.

Optional

Zhang et al., GSM1k (2405.00332). Bhagia et al., Model Ladders (2412.04403). Chiang et al., Chatbot Arena (2403.04132). Magnusson et al., Paloma (2312.10523).