# **Evaluation** CSC2541H1 Topics in Machine Learning, Winter 2025, UToronto

Chris J. Maddison

# Announcements

- after class.
- This week we again only have two presentations.

**Questions?** 

### • If you are assigned to present on March 21, come talk about presentations

# Today

- models.
- Today we're going to talk about evaluation.

### • We've now basically covered the full training pipeline of contemporary large

# Why evaluate?

- Benchmarks drive research progress and industry adoption
- Evals incentivize the research community and dictate optimization targets
- Evals help us detect progress (improvements in good capabilities) or threats (safety concerns)



### Open LLM Leaderboard

	Rank	Туре	Model	Average ① 0	IFEval ① ≎	BE
푸	1	<b></b>	MaziyarPanahi/calme-3.2-instruct-78b 🛛	• 52.08 %	80.63 %	62
꾸	2	9	MaziyarPanahi/calme-3.1-instruct-78b 🛛	• 51.29 %	81.36 %	62
꾸	3	9	dfurman/CalmeRys-78B-Orpo-v0.1 🛛	• 51.23 %	81.63 %	61
꾸	4	9	MaziyarPanahi/calme-2.4-rys-78b 🛛	• 50.77 %	80.11 %	62
푸	5	<b></b>	huihui-ai/Qwen2.5-72B-Instruct-abliterated 🛛	• 48.11 %	85.93 %	60
꾸	6	9	Qwen/Qwen2.5-72B-Instruct	• 47.98 %	86.38 %	61
꾸	7	9	MaziyarPanahi/calme-2.1-qwen2.5-72b 🛛	• 47.86 %	86.62 %	61
꾸	8	<b></b>	newsbang/Homer-v1.0-Qwen2.5-72B 🛛	• 47.46 %	76.28 %	62
꾸	9	9	ehristoforu/qwen2.5-test-32b-it 🛛	• 47.37 %	78.89 %	58
푸	10	<b></b>	Saxo/Linkbricks-Horizon-AI-Avengers-V1-32B	• 47.34 %	79.72 %	57



## Types of benchmarks

- Static vs Dynamic
- Preference-based vs Ground Truth

### **Question Source**

		Static	Live
Evaluation	Ground Truth	MMLU, HellaSwag, GSM-8K	Codeforces Weekly Conte
Metric	Human Preference	MT-Bench, AlpacaEval	LLM Arena

Chiang et al, 2024, Chatbot Arena



## GSM8K Static, Ground-Truth

problems created by human problem writers

**Problem:** Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?

**Solution:** Beth bakes 4 2 dozen batches of cookies for a total of 4\*2 = <<4\*2=8>>8 dozen cookies There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12\*8 = <<12\*8=96>>96 cookies She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies Final Answer: 6

# Dataset of 8.5K high quality linguistically diverse grade school math word

Cobbe et al, 2021, Training Verifers to Solve Math Word Problems



## **GSN8K** Static, Ground-Truth

- Injected calculation annotations into train set.
- At test-time, a calculator will override sampling when the model chooses to use these annotations.
  - Calculator override reduces arithmetic errors. Is this fair?
- Typically disabled now.



Cobbe et al, 2021, Training Verifers to Solve Math Word Problems

## **Chatbot Arena** Dynamic, Preference-based

- Open platform for anonymous, randomized LLM battles
- Pairwise comparisons with hidden model identities

### C Expand to see 20+ Arena players

### 🗇 Model A

I would like to buy a new car. Can you give me three recommendations? A family-friendly car with lots of space, An eco-friendly electric or hybrid car. Family outings or road trips, City driving and errands, Long distance travels. Give me three options at different price points.

Certainly! When looking for a new car that fits the criteria of being familyfriendly with lots of space, eco-friendly, and suitable for various driving needs, you'll want to consider factors like safety features, cargo space, seating comfort, fuel efficiency, and range. Here are three options at different price points:

1. Budget-Friendly Option: Toyota RAV4 Hybrid

 The Toyota RAV4 Hybrid is a great option for families looking for an affordable eco-friendly vehicle. It offers ample space for passengers and cargo, making it suitable for family outings and road trips. It's also efficient for city driving and errands due to its hybrid powertrain, which combines a gasoline engine with an electric motor to improve fuel economy. The RAV4 Hybrid provides safety features and a comfortable ride, making it a great

👈 A is better

👉 B is better

### 🗊 Model B

I would like to buy a new car. Can you give me three recommendations? A family-friendly car with lots of space, An eco-friendly electric or hybrid car. Family outings or road trips, City driving and errands, Long distance travels. Give me three options at different price points.

Absolutely! Here are three car options for your diverse needs, each aiming for a different budget:

### Family-Friendly, Spacious:

• Chrysler Pacifica Hybrid (Starting at ~\$40,000): This minivan is the pinnacle of family-friendly transportation. It offers abundant space for passengers and cargo (up to 140.5 cubic feet with seats folded!), plus clever features like Stow 'n Go second-row seats and built-in vacuum cleaner. The hybrid powertrain delivers 36 mpg combined, reducing environmental impact and fuel costs. It's perfect for road trips, city errands, and long-distance travels thanks to its comfortable ride and long driving range.

### **Eco-Friendly Electric:**

🔝 Tie

👎 Both are bad

Send

Enter your prompt and press ENTER

Chiang et al, 2024, Chatbot Arena



## **Chatbot Arena Dynamic, Preference-based**

modelled as (where  $s(m) \in \mathbb{R}$  is a score associated with m)

$$P(H=1) = -\frac{1}{1}$$

- Score function is estimated by a cross-entropy objective over all pairwise preferences.
- The ranking is then (roughly)

 $\operatorname{rank}(m) = 1 + 1$ 

and management of uncertainty.

• Model preferences in a Bradley-Terry model, i.e. the probability that a human prefers m to m' is

 $+ \exp(s(m') - s(m))$ 

$$\sum_{m'} 1\{s(m') > s(m)\}$$

I'm omitting some important details that correct for non-uniform choices of which models to serve

# Challenges

- Static eval challenges:
  - saturation, contamination, scalability
- Preference-based challenges:
  - subjectivity, consistency, reward hacking

### Static evals have dynamic ranges



### Preference evals lead to reward hacking



Rob Mulla



## **Good benchmarks** Jason Wei's Blogpost

- 1. Many examples
- 2. High quality, low noise
- 3. One metric
- 4. Simple to run
- 5. A meaningful task
- 6. Correct grading
- 7. Large dynamic range

### Static benchmarks saturate over time



### Noise reduces the dynamic range



Vendrow et al, 2025, Do Large Language Model Benchmarks Test Reliability?

## **Automated evaluation** AlpacaFarm

- Using stronger models (like) GPT-4) to approximate human judgment
- Benefits: scalability, consistency, cost-effectiveness
- Limitations: judge model bias, risk of reinforcing limitations



\$70

\$3,150

Hours

Days

×



Ъ.

Alpaca Farm

Human Feedback

# Urgent challenges

- How can we evaluate capabilities that are beyond human capabilities?
- Safety benchmarking (next week) is also a critical issues.
  - How can we measure capabilities relevant to catastrophic risk if we're still researching the contours of those outcomes?
  - Ideally, we'd rule-in risks.

## **Take-homes** Evaluation

- Benchmarking drive progress
- Benchmark development is very high impact
- Benchmark development is challenging