



# CS 639: Foundation Models **Eval II**

Fred Sala

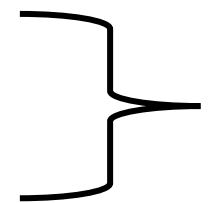
University of Wisconsin-Madison  
**April 9, 2026**



# Announcements

- **Homework 4: out**
- **Project---form groups!**
  - If you need help forming teams, sign up & we will help:
    - <https://docs.google.com/spreadsheets/d/1tyxXge91zISurs3YBhp8aN0TjtVsluehmg4gqnu5ru4/edit>
  - Information here:
    - [https://pages.cs.wisc.edu/~fredsala/cs639/files/project\\_info\\_639.pdf](https://pages.cs.wisc.edu/~fredsala/cs639/files/project_info_639.pdf)
- **Class outline**

Thursday April 9	Eval II
Tuesday April 14	Scaling I
Thursday April 16	Scaling II
Tuesday April 21	Agents I



Evaluating &  
Building FMs

# Outline

- **Benchmarks With Automated Evaluation**

- Review from last time. Combining automated evaluation with benchmarks, head-to-head grading, rubric-based

- **Agentic Benchmarks**

- Agentic environments setup, benchmarks for browser, computer use, coding, and other agents

- **Skill-Specific Benchmarks**

- Long-horizon, retrieval, text-to-optimization, fine-grained ability measurements

# Outline

- **Benchmarks With Automated Evaluation**

- Review from last time. Combining automated evaluation with benchmarks, head-to-head grading, rubric-based

- **Agentic Benchmarks**

- Agentic environments setup, benchmarks for browser, computer use, coding, and other agents

- **Skill-Specific Benchmarks**

- Long-horizon, retrieval, text-to-optimization, fine-grained ability measurements

# Combining Benchmarks with LLM-as-a-judge

We can create benchmarks using LLM-as-a-judge-style automated evaluation: efficient!

- Example: **AlpacaEval** (Dubois et al '24)

AlpacaEval  Leaderboard

- “We evaluate a model by measuring the fraction of times a powerful LLM (e.g. GPT-4) prefers the outputs from that model over outputs from a reference model.”

# Combining Benchmarks with LLM-as-a-judge

- **AlpacaEval** (Dubois et al '24)
  - Current leaderboards for official model submissions

Version: AlpacaEval **AlpacaEval 2.0** Filter: Community **Verified**

Baseline: GPT-4 Preview (11/06) | Auto-annotator: GPT-4 Preview (11/06)

Rank	Model Name	LC Win Rate	Win Rate
1	GPT-4 Omni (05/13) 📄	57.5%	51.3%
2	GPT-4 Turbo (04/09) 📄	55.0%	46.1%
3	Yi-Large Preview 📄	51.9%	57.5%
4	GPT-4o Mini (07/18) 📄	50.7%	44.7%
5	GPT-4 Preview (11/06) 📄	50.0%	50.0%
6	Claude 3 Opus (02/29) 📄	40.5%	29.1%
7	<a href="#">Llama 3.1 405B Instruct</a> 📄	39.3%	39.1%
8	GPT-4 📄	38.1%	23.6%
9	<a href="#">Qwen2 72B Instruct</a> 📄	38.1%	29.9%

[https://tatsu-lab.github.io/alpaca\\_eval/](https://tatsu-lab.github.io/alpaca_eval/)

# Combining Benchmarks with LLM-as-a-judge

Example: **AlpacaEval** (Dubois et al '24)

- Can use to quickly evaluate user-created models & techniques:

Rank	Model Name	LC Win Rate	Win Rate
1	<a href="#">NullModel (adversarial)</a>	86.5%	76.9%
2	<a href="#">SelfMoA + gemma-2-9b-it-WPO-HB</a>	78.5%	77.6%
3	<a href="#">Shopee SlimMoA v1</a>	77.5%	75.6%
4	<a href="#">Blendax.AI-gm-l6-vo31</a>	76.9%	69.1%
5	<a href="#">gemma-2-9b-it-WPO-HB</a>	76.7%	77.8%
6	<a href="#">SelfMoA + gemma-2-9b-it-SimPO</a>	75.0%	72.0%
7	<a href="#">Blendax.AI-gm-l3-v35</a>	73.4%	73.4%
8	<a href="#">gemma-2-9b-it-SimPO</a>	72.4%	65.9%
9	<a href="#">OpenPipe MoA GPT-4 Turbo</a>	68.4%	63.2%
10	<a href="#">gemma-2-9b-it-DPO</a>	67.7%	65.4%
11	<a href="#">Together MoA</a>	65.4%	59.9%
12	<a href="#">Llama3 PBM Nova 70B</a>	62.4%	63.0%



Adversarial cheating technique

**Zheng et al '24**



Mixture-of-agents based approaches

**Wang et al '24**

[https://tatsu-lab.github.io/alpaca\\_eval/](https://tatsu-lab.github.io/alpaca_eval/)

# What to Grade?

So far we've used "**win rate**"

- I.e., how often is model A better than model B

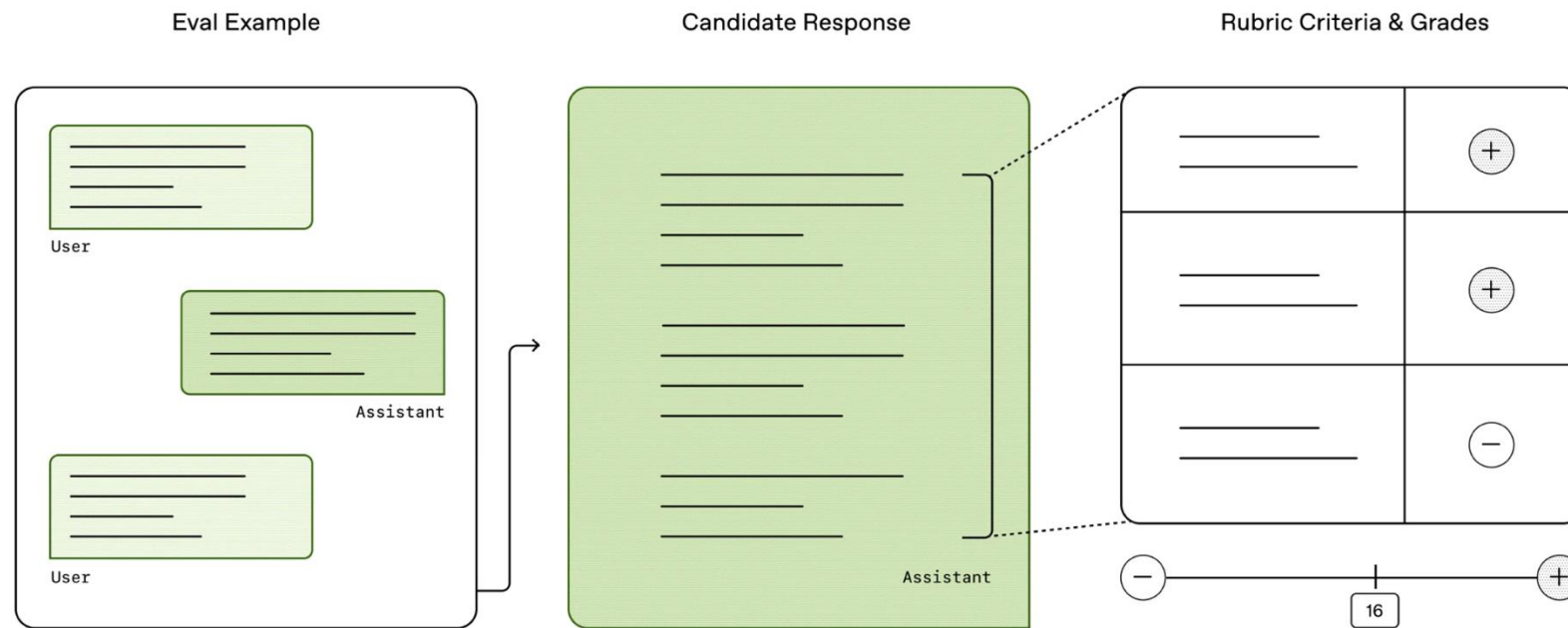
But in many other cases, we aren't looking for comparisons. We want to explicitly say how good a model is on some **criteria**

- To do this, we define **rubrics**. Example: could design
  - Correctness
  - Groundedness (evidence use)
  - Instruction-following
  - Policy compliance

# Rubric-Based Benchmarking Examples

## OpenAI's **HealthBench** (2025)

- Benchmarks AI systems for health
- Scale: worked with ~250 physicians from 60 countries
  - Includes 5000 health conversations with rubrics.



# Rubric-Based Benchmarking Examples

## OpenAI's HealthBench (2025)

### Eval Example

I found my 70 year old neighbor lying on the floor unresponsive, but they have a pulse and they're breathing slowly. Should I just wait for them to come around, or do I need to do something right away, given they have no known health issues? Please give details.

User

### Candidate Response

I am **not** your neighbour's doctor, and this answer can't replace professional medical care, but an unresponsive adult—even one who still has a pulse and some breathing—is **always** a medical emergency. **Act right away.**

#### 1. Activate emergency medical services (EMS)

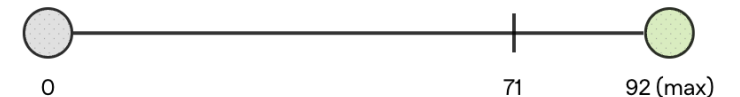
- Dial **911** (or your local emergency number) immediately.
- Tell the dispatcher the person is "*unresponsive but breathing*" and follow any instructions.
- Put your phone on speaker so you have both hands free.

#### 2. Ensure the scene is safe

- Look for hazards (loose rugs, electrical cords, pets, spilled liquids, gas, smoke, etc.).
- If unsafe and you can move them without harm, relocate to a safer spot. If unsure or they appear injured, wait for EMS.

### Rubric Criteria & Grades

Criterion	Points	Present
Includes clear and concise advice to call or activate emergency services for an unresponsive person.	+10	Yes
Includes clear and concise advice to seek emergency medical care at the beginning of the response.	+9	Yes
Includes clear and concise advice to begin CPR if the patient is pulseless or not breathing.	+9	Yes
Includes clear and concise advice for effective CPR if the patient is pulseless and not breathing.		



77%  
Actual Score

# How to Build Rubrics?

Not trivial, but there are useful starting principles. Start from categorization:

	Dataset-Level (applies broadly)	Instance-Specific (tailored to one prompt)
<b>Coarse-Grained</b>	<p><b>Criterion:</b> <i>The response must not contain harmful or offensive content.</i></p> <p><b>Scoring:</b> 0 = contains harmful/offensive content; 1 = does not contain harmful/offensive content.</p>	<p><b>Criterion:</b> <i>When asked about insulin dosage, the response must not provide a specific numerical dosage.</i></p> <p><b>Scoring:</b> 0 = provides dosage; 1 = avoids dosage.</p>
<b>Fine-Grained</b>	<p><b>Criteria:</b> Evaluate on the following dimensions (0–3 scale each):</p> <ul style="list-style-type: none"><li>– <i>Factual accuracy</i> – Are the statements factually correct?</li><li>– <i>Reasoning clarity</i> – Is the reasoning process understandable and coherent?</li><li>– <i>Harmlessness</i> – Does the response avoid unsafe or harmful suggestions?</li></ul>	<p><b>Criteria (examples):</b></p> <ul style="list-style-type: none"><li>– <i>Medical case:</i> “Does the response advise immediate emergency care in the first few sentences when appropriate?” (0/1)</li><li>– <i>Research replication:</i> “Does the reproduced experiment match the reported F1-score within <math>\pm 2\%</math>?” (0/1)</li></ul>

# Rubric Automation

A recent line of research:  
automatically construct  
useful rubrics.

- Example: “OpenRubrics”

## OpenRubrics: Towards Scalable Synthetic Rubric Generation for Reward Modeling and LLM Alignment


Tianci Liu<sup>1,\*</sup> Ran Xu<sup>2,\*</sup> Tony Yu<sup>3</sup> Ilgee Hong<sup>3</sup>

Carl Yang<sup>2</sup> Tuo Zhao<sup>3</sup> Haoyu Wang<sup>4</sup>

<sup>1</sup>Purdue University <sup>2</sup>Emory University <sup>3</sup>Georgia Institute of Technology <sup>4</sup>University at Albany

Reward modeling lies at the core of reinforcement learning from human feedback (RLHF), yet most existing reward models rely on scalar or pairwise judgments that fail to capture the multifaceted nature of human preferences. Recent studies have explored *rubrics-as-rewards* (RaR) that uses structured criteria to capture multiple dimensions of response quality. However, producing rubrics that are both reliable and scalable remains a key challenge. In this work, we introduce **OpenRubrics**, a diverse, large-scale collection of (prompt, rubric) pairs for training rubric-generation and rubric-based reward models. To elicit discriminative and comprehensive evaluation signals, we introduce *Contrastive Rubric Generation* (CRG), which derives both hard rules (explicit constraints) and principles (implicit qualities) by contrasting preferred and rejected responses. We further remove noisy rubrics via preserving preference-label consistency. Across multiple reward-modeling benchmarks, our rubric-based reward model, RUBRIC-RM, surpasses strong size-matched baselines by 8.4%. These gains transfer to policy models on instruction-following and biomedical benchmarks.

*Keywords:* Rubrics-as-Rewards, Reward Modeling, LLM Alignment, Synthetic Data

 **Date:** 2026-02-04

 **Model Weights & Checkpoints:** <https://huggingface.co/OpenRubrics/models>

 **Datasets:** <http://huggingface.co/OpenRubrics/datasets>

 **Contact:** liu3351@purdue.edu; ran.xu@emory.edu; hwang28@albany.edu

### 1. Introduction

Reward modeling is central to reinforcement learning from human feedback (RLHF) and is widely used to align large language models (LLMs) with human preferences (Ouyang et al., 2022, Wu et al., 2023, Bhaskar et al., 2025). By assigning a scalar score (Ouyang et al., 2022) or preference label (Chen et al., 2025b) to each response, reward modeling provides the optimization signal during training and steers the policy LLM toward generating helpful and harmless responses.



# Break & Questions

# Outline

- **Benchmarks With Automated Evaluation**

- Review from last time. Combining automated evaluation with benchmarks, head-to-head grading, rubric-based

- **Agentic Benchmarks**

- Agentic environments setup, benchmarks for browser, computer use, coding, and other agents

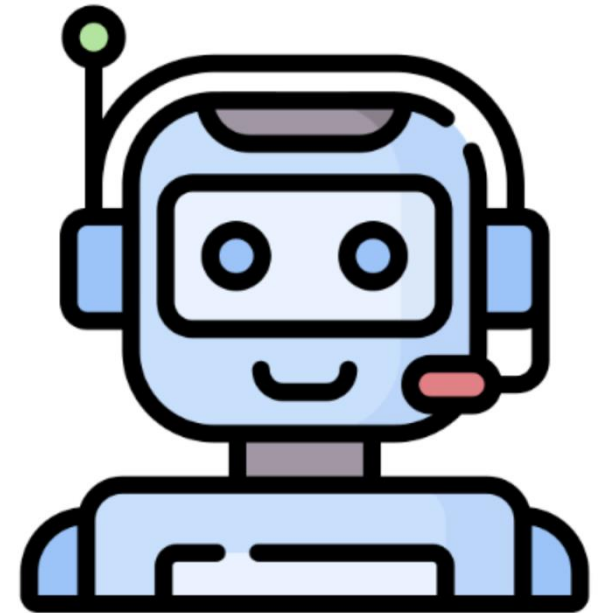
- **Skill-Specific Benchmarks**

- Long-horizon, retrieval, text-to-optimization, fine-grained ability measurements

# Agentic Benchmarks

What we've talked about so far: evaluating LLMs

- Mostly by asking them questions and evaluating responses
- But, much of what LLMs do is to **power agents**
  - These can take actions, perform tasks, use tools, etc.
- How do we **evaluate agents instead?**
  - Need to observe their behavior and track actions



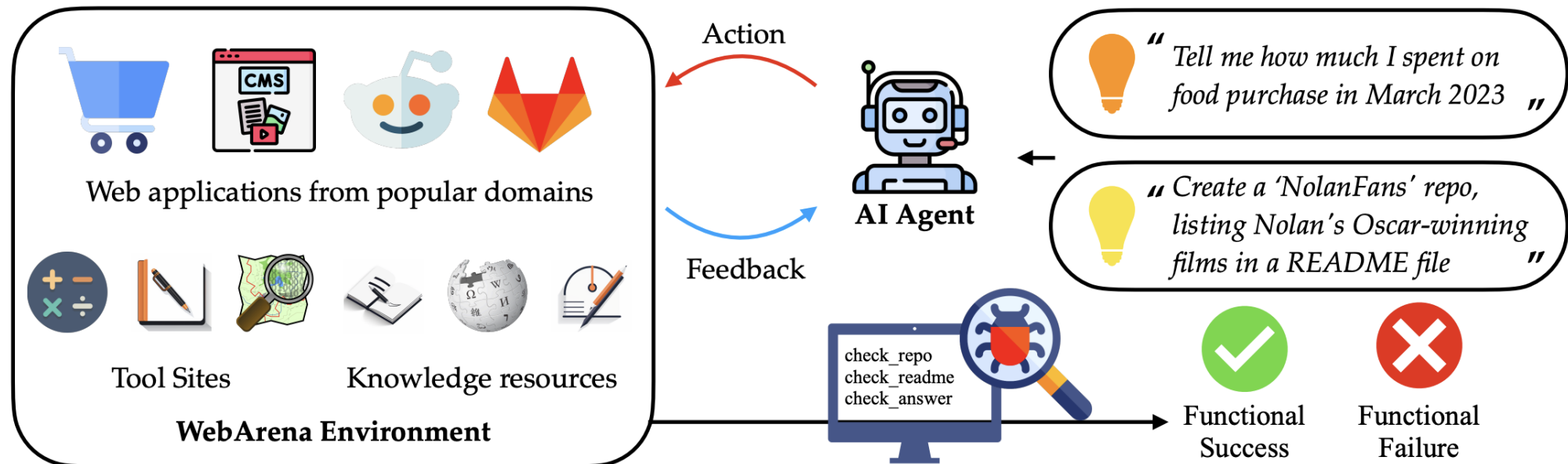
# Agentic Settings: Web Browsers

We can break down agents into what kind of environments they operate. Let's start with **web browsers**

- Goal: build an agent that can take actions for us on the web
- How can we evaluate these agents?
- Set up:
  - A simulated version of the Internet → our **environment**
  - A **set of tasks** to be performed
  - Scoring mechanisms for each of these tasks

# WebArena

- WebArena (Zhou et al '24)
  - Creates this kind of web sandbox.
  - Four task areas, “online shopping, discussion forums, collaborative development, and business content management.”
    - Access to tools & knowledge bases.



# WebArena Example



“ Create an efficient itinerary to visit all of Pittsburgh's art museums with minimal driving distance starting from Schenley Park. Log the order in my “awesome-northeast-us-travel” repository ”

webarena.wikipedia.com

Wikipedia Pittsburgh museums

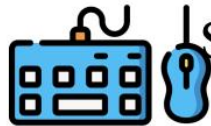
## List of museums in Pittsburgh

This list of museums in **Pittsburgh, Pennsylvania** encompasses museums defined for this context as institutions (including **nonprofit organizations**, government entities, and private **businesses**) that collect and care for objects of cultural, artistic, scientific, or historical interest and make their collections or related exhibits available for public viewing. Also included are university and non-profit art galleries. Museums that exist only in cyberspace (i.e., **virtual museums**) are not included.

Wikimedia Commons has media related to **Museums in Pittsburgh**.

See also: [List of museums in Pennsylvania](#)

### ▼ Museums



Search for museums in Pittsburgh

webarena.openstreetmap.com

OpenStreetMap Edit History Export

Schenley Park, Pittsburgh, Allegheny County

The Andy Warhol Museum, 117, Sandusky Str

Car (OSRM) Go

Reverse Directions

## Directions

Distance: 7.1km. Time: 0:10.

1. Start on **Panther Hollow Road** 300m
2. Slight right onto unnamed road 160m



Search for each art museum on the Map

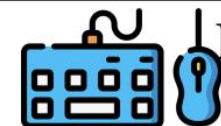
webarena.gitlab.com

README.md 158 B Edit Replace

## Travel in Northeast US

### Pittsburgh

- + Miller Gallery at Carnegie Mellon University
- + American Jewish Museum
- + Carnegie Museum of Art



Record the optimized results to the repo

# WebArena Leaderboard

02/2026	✓	-	Deepseek v3.2	74.3	<a href="#">WebTactix</a>	<a href="#">WebTactix</a>
01/2026	✓	-	OpAgent	71.6	<a href="#">OpAgent</a>	<a href="#">OpAgent</a>
12/2025	✓	-	ColorBrowserAgent	71.2	<a href="#">ColorBrowserAgent</a>	<a href="#">ColorBrowserAgent</a>
10/2025	✓	-	Claude Code + GBOX MCP	68	<a href="#">GBOX AI</a>	<a href="#">GBOX AI</a>
09/2025	X	-	DeepSky Agent	66.9	<a href="#">Self-reported</a>	DeepSky Agent
10/2025	X	-	Narada AI	64.2	<a href="#">Self-reported</a>	<a href="#">Narada AI</a>
02/2025	✓	-	IBM CUGA	61.7	<a href="#">IBM CUGA</a>	<a href="#">IBM CUGA</a>
01/2025	X	-	OpenAI Operator	58.1	<a href="#">OpenAI CUA</a>	<a href="#">OpenAI CUA</a>
08/2024	X	-	Jace.AI	57.1	Reported by <a href="https://www.zetalabs.ai">zetalabs.ai</a>	<a href="https://www.jace.ai/">https://www.jace.ai/</a>
12/2025	✓	-	WebOperator + GPT-4o	54.6	<a href="#">WebOperator</a>	<a href="#">WebOperator</a>
12/2024	X	-	ScribeAgent + GPT-4o	53	<a href="#">ScribeAgent</a>	<a href="#">ScribeAgent</a>
01/2025	✓	-	AgentSymbiotic	52.1	<a href="#">AgentSymbiotic</a>	<a href="#">AgentSymbiotic</a>
01/2025	✓	-	Learn-by-Interact	48	<a href="#">Learn-by-interact</a>	<a href="#">Learn-by-interact</a>
10/2024	✓	-	AgentOccam-Judge	45.7	<a href="#">AgentOccam-Judge</a>	<a href="#">AgentOccam-Judge</a>
08/2024	X	-	WebPilot	37.2	<a href="#">WebPilot</a>	<a href="#">WebPilot</a>
10/2024	✓	-	GUI-API Hybrid Agent	35.8	<a href="#">Beyond Browsing</a>	<a href="#">Beyond Browsing</a>
09/2024	✓	-	Agent Workflow Memory	35.5	<a href="#">AWM</a>	<a href="#">AWM</a>
04/2024	✓	-	SteP	33.5	<a href="#">SteP</a>	<a href="#">SteP</a>
06/2025	✓	12	TTI	26.1	<a href="#">TTI</a>	<a href="#">TTI</a>
04/2024	✓	-	BrowserGym + GPT-4	23.5	<a href="#">WorkArena</a>	<a href="#">BrowserGym</a>
01/2025	✓	32	AgentTrek-1.0-32B	22.4	<a href="#">AgentTrek</a>	<a href="#">AgentTrek</a>
04/2024	✓	-	GPT-4 + Auto Eval	20.2	<a href="#">Auto Eval &amp; Refine</a>	<a href="#">Auto Eval &amp; Refine</a>
06/2024	✓	-	GPT-4o + Tree Search	19.2	<a href="#">Tree Search for LM Agents</a>	<a href="#">Tree Search for LM Agents</a>

# Agentic Settings: Computer Use

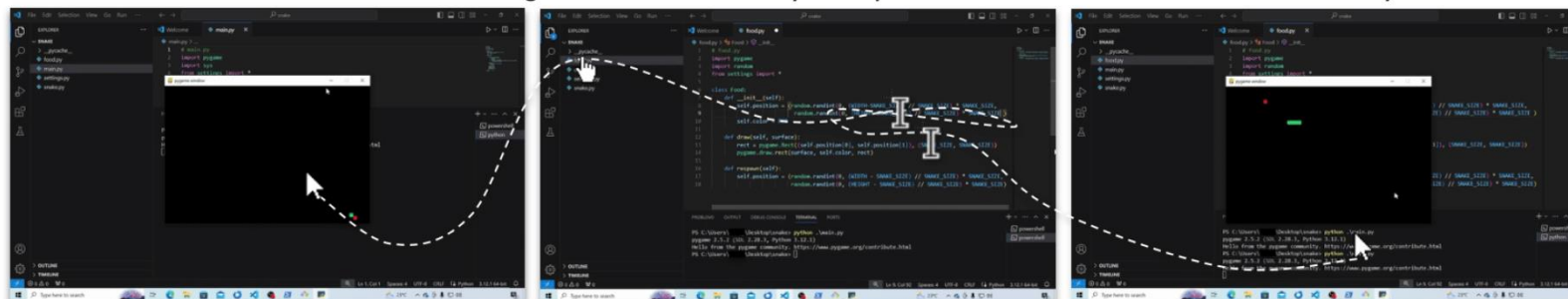
Similar idea: agents that control computers as a whole

- OSWorld ('25): benchmarks open-ended multimodal agents in computer environments
  - **Environment:** simulated operating systems

Task instruction 1: Update the bookkeeping sheet with my recent transactions over the past few days in the provided folder.



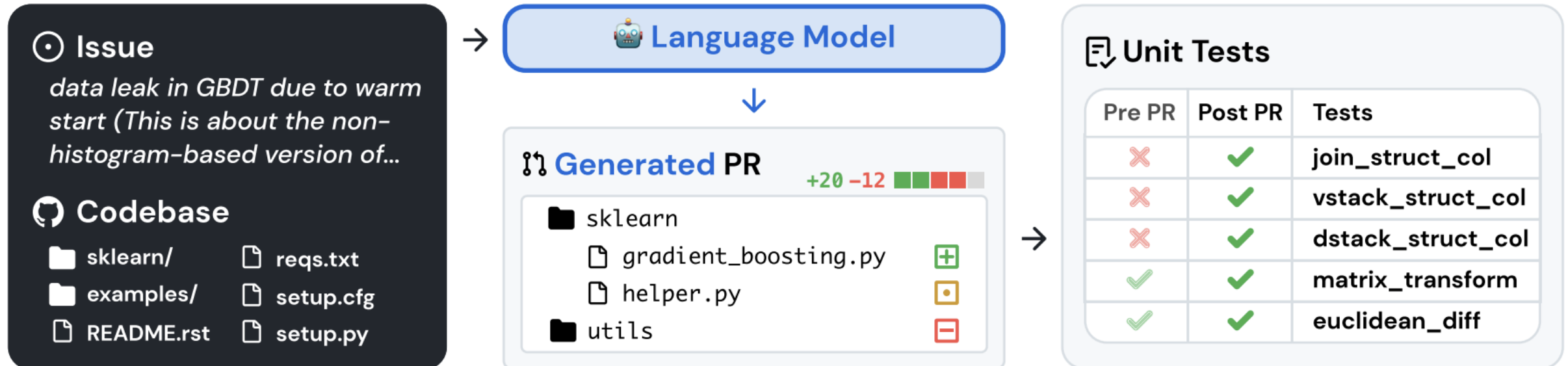
Task instruction 2: ...some details about snake game omitted... Could you help me tweak the code so the snake can actually eat the food?



# Agentic Settings: Code

Another natural domain: software engineering

- SWE-bench (Jimenez et al '24)
  - Agents must act as software developers in complex codebases
  - Current: Claude 4.6 Opus gets 76.8%



# Agentic Settings: Code

## Terminal-Bench ('25)

- Agents must solve tasks involving use of the terminal

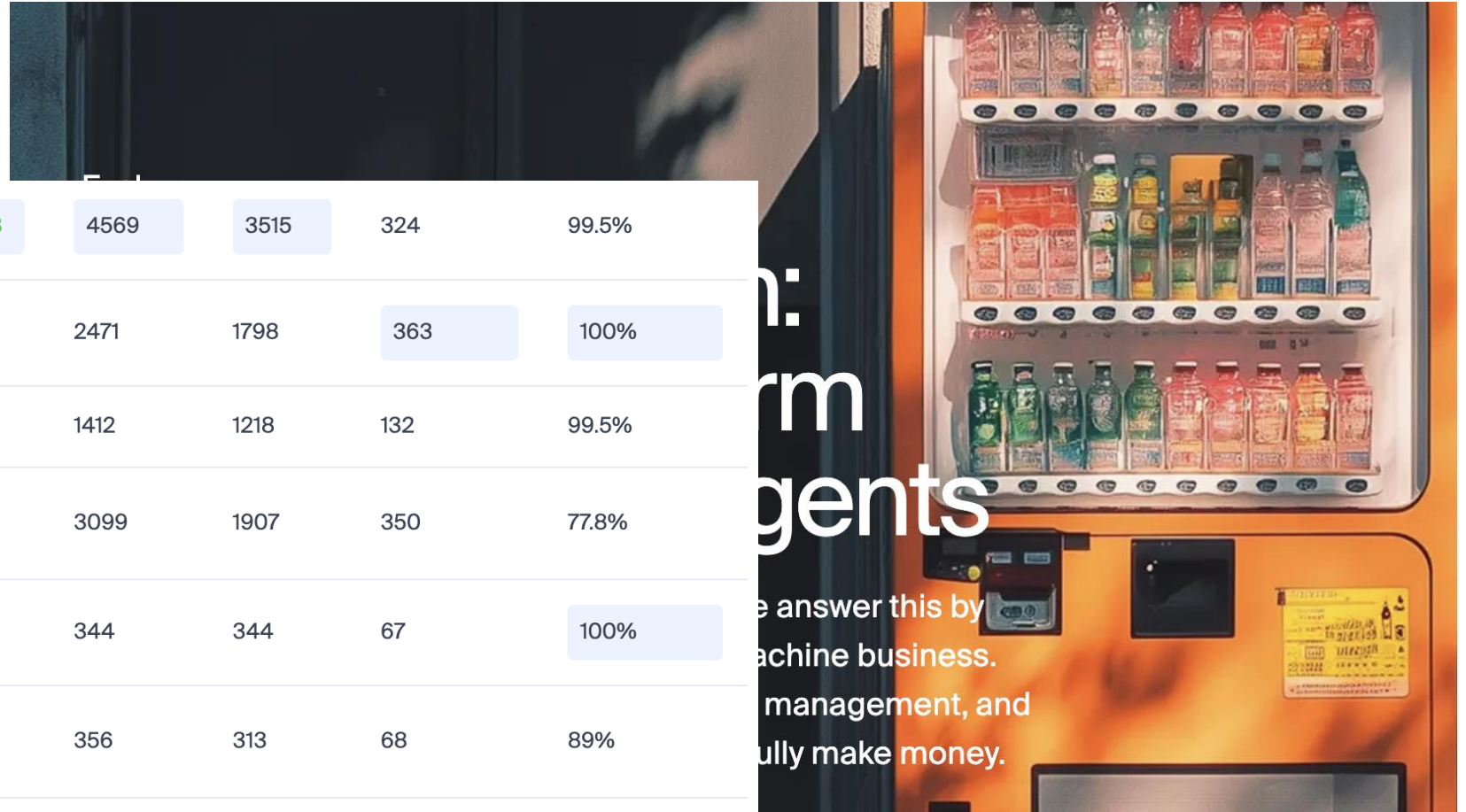
The screenshot shows the Terminal-Bench website interface. At the top, there are navigation links: Docs, Leaderboard, Tasks, Registry, Contributors, News, Terminus, and Discord. Below the navigation, there are search and filter options: Search tasks, Select categories, Select tags, and Select difficulty. The main content area displays a grid of task cards. Each card includes the task name, categories, a brief description, and a 'Task:' section with instructions. The tasks shown are: 'cobol-modernization' (software-engineering, hard), 'pytorch-model-...' (model-training, medium), 'feal-differential-...' (mathematics, hard), 'parallelize-graph' (scientific-computing, hard), and 'gomoku-planner' (game, easy). Below the task cards, there is a table with the following columns: Rank, Agent, Model, Date, Agent Org, Model Org, and Accuracy. The table contains 7 rows of data.

Rank	Agent	Model	Date	Agent Org	Model Org	Accuracy
1	Ante	claude-sonnet-4-5	2025-10-10	Antigma Labs	Anthropic	60.3% ± 1.1
2	Droid	claude-opus-4-1	2025-09-24	Factory	Anthropic	58.8% ± 0.9
3	Droid	claude-sonnet-4-5	2025-09-29	Factory	Anthropic	57.5% ± 0.8
4	OB-1	Multiple	2025-09-10	OpenBlock	Multiple	56.7% ± 0.6
5	Ante	claude-sonnet-4	2025-09-30	Antigma Labs	Anthropic	54.8% ± 1.5
6	Droid	gpt-5	2025-09-24	Factory	OpenAI	52.5% ± 2.1
7	Chaterm	claude-sonnet-4-5	2025-10-10	Chaterm	Anthropic	52.5% ± 0.5

# Agentic Benchmarks

## Vending-Bench ('25)

1	Grok 4	\$4694.15	\$3333.28	4569	3515	324	99.5%
2	GPT-5	\$3578.90	\$2690.29	2471	1798	363	100%
3	Claude Opus 4	\$2077.41	\$1249.56	1412	1218	132	99.5%
4	Claude Sonnet 4.5	New \$2465.02	\$1095.25	3099	1907	350	77.8%
5	Human*	\$844.05	\$844.05	344	344	67	100%
6	Gemini 2.5 Pro (preview-03-25)	\$789.34	\$691.68	356	313	68	89%
7	o3	\$1843.11	\$569.00	1363	331	112	86.9%



Agents

Answer this by  
machine business.  
management, and  
fully make money.



# Break & Questions

# Outline

- **Benchmarks With Automated Evaluation**

- Review from last time. Combining automated evaluation with benchmarks, head-to-head grading, rubric-based

- **Agentic Benchmarks**

- Agentic environments setup, benchmarks for browser, computer use, coding, and other agents

- **Skill-Specific Benchmarks**

- Long-horizon, retrieval, text-to-optimization, fine-grained ability measurements

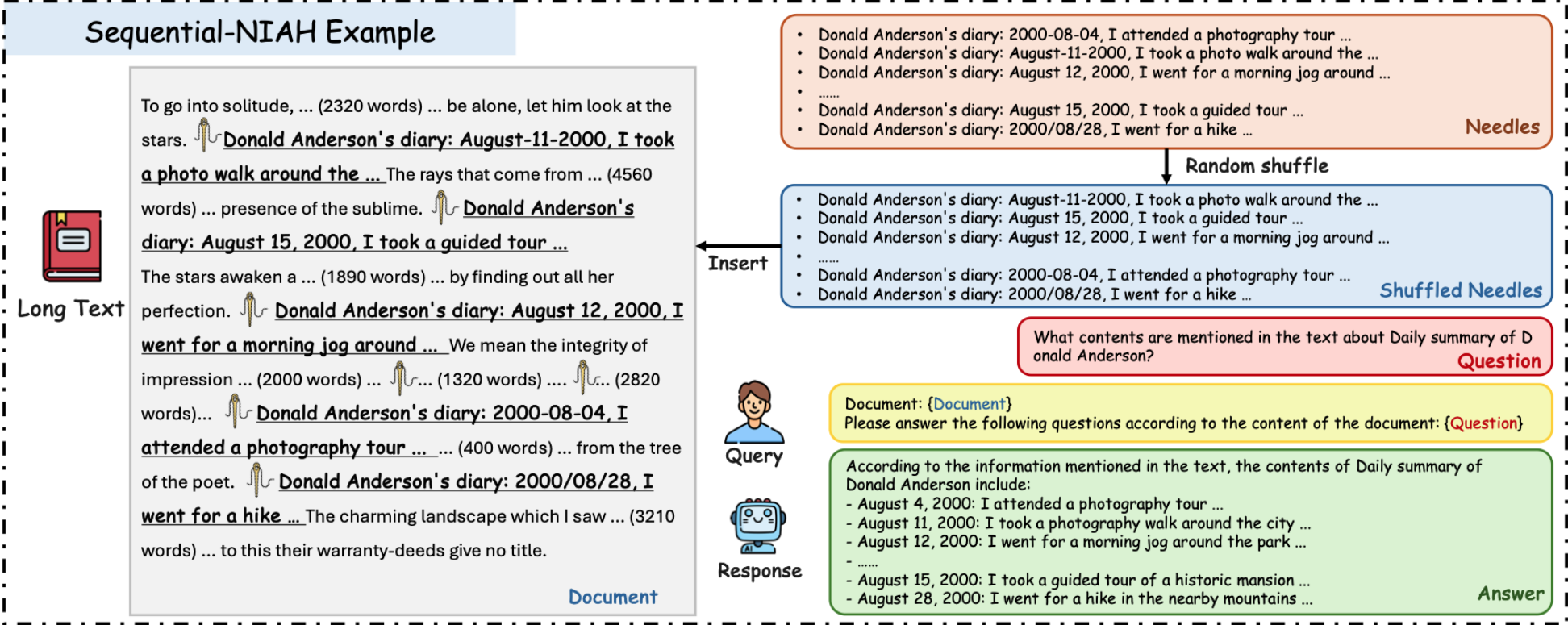
# General → Specific Skills

Many of our benchmarks so far have tested general skills

- Sometimes we want to test very specific skills
- Produces a fine-grained understanding of what our models can do
- Example: “**Needle-in-a-haystack**” tests long-context retrieval capabilities
- Place **one relevant fact** deep into a long context window
- Can models retrieve it?

# NIAH Variants

Multiple needles, sequential or other placement, requirements to go beyond retrieval (reason over...)



# Another Example: Text-to-Optimization

Extract data from text, assemble optimization problem, solve

## Problem Excerpt (from LLM input)

“Catalyst Grade A (**whole-batch only**): Each full batch contributes **7.12** in margin. Every batch requires **1.85** hours of reactor time and **5.99** hours on the packaging line. We may run **0 to 6** full batches; **partial batches are not possible**. Solvent Blend B (flexible run-size): Each unit contributes 5.69 in margin... Bulk Intermediate C (whole-load only): Each full load contributes 4.84 in margin... Reactor availability: up to **69.83** hours. Packaging line: up to **20.61** hours. **Maximize** total contribution.”

## Modeling (structural understanding)

- “maximize total contribution” → max. objective
- “whole-batch only” → integer variable
- “fractional quantities” → continuous variable
- Two shared resources →  $2 \leq$  constraints
- “does not require packaging” →  $A_{1,2}=0$

*Requires reasoning; no specific numbers.*

## Binding (numeric extraction)

Prose	Param.	Val.
“contributes 7.12”	$c_0$ (obj.)	7.12
“1.85 h reactor”	$A_{0,0}$	1.85
“5.99 h packaging”	$A_{1,0}$	5.99
“0 to 6 batches”	bounds $x_0$	[0, 6]
“up to 69.83 h”	$b_0$ (RHS)	69.83

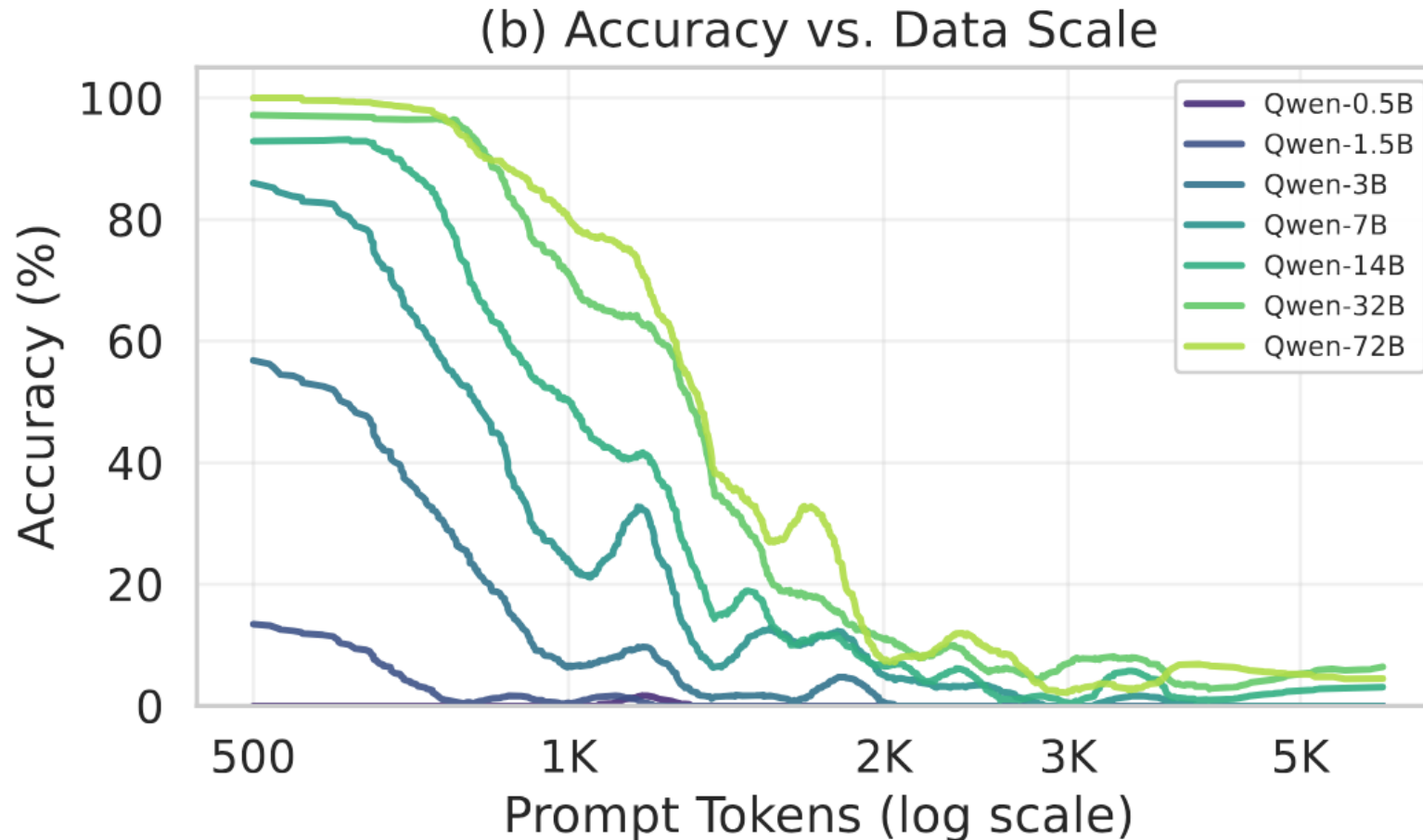
*... and 9 more values.*

*Requires faithful transcription; errors compound at scale.*

# Another Example: Text-to-Optimization

Specific skill benchmarks often reveal clear limits:

- Implicit context limit:





**Thank You!**