



COMP 336 | Natural Language Processing

Lecture 18: LLM/VLM benchmarks and evaluation

Spring 2025

Evaluating natural language generation

Types of text evaluation methods

Ref: They walked **to the grocery store.**
Gen: **The woman went to the hardware store.**

Content Overlap Metrics



Model-based Metrics

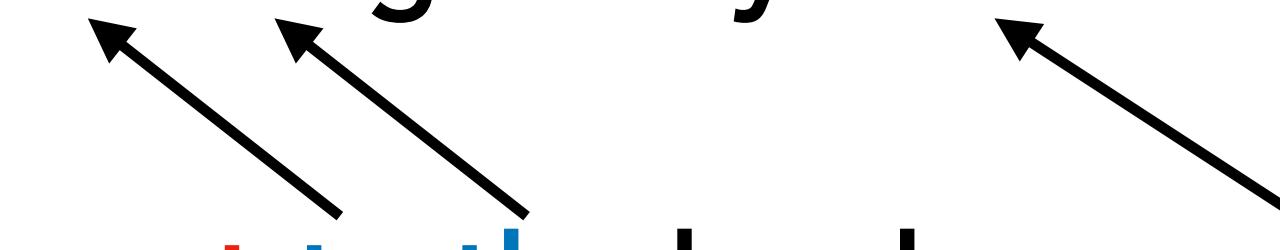


Human Evaluation

Content overlap metrics

Ref: They walked **to the grocery store.**

Gen: **The woman went to the hardware store.**



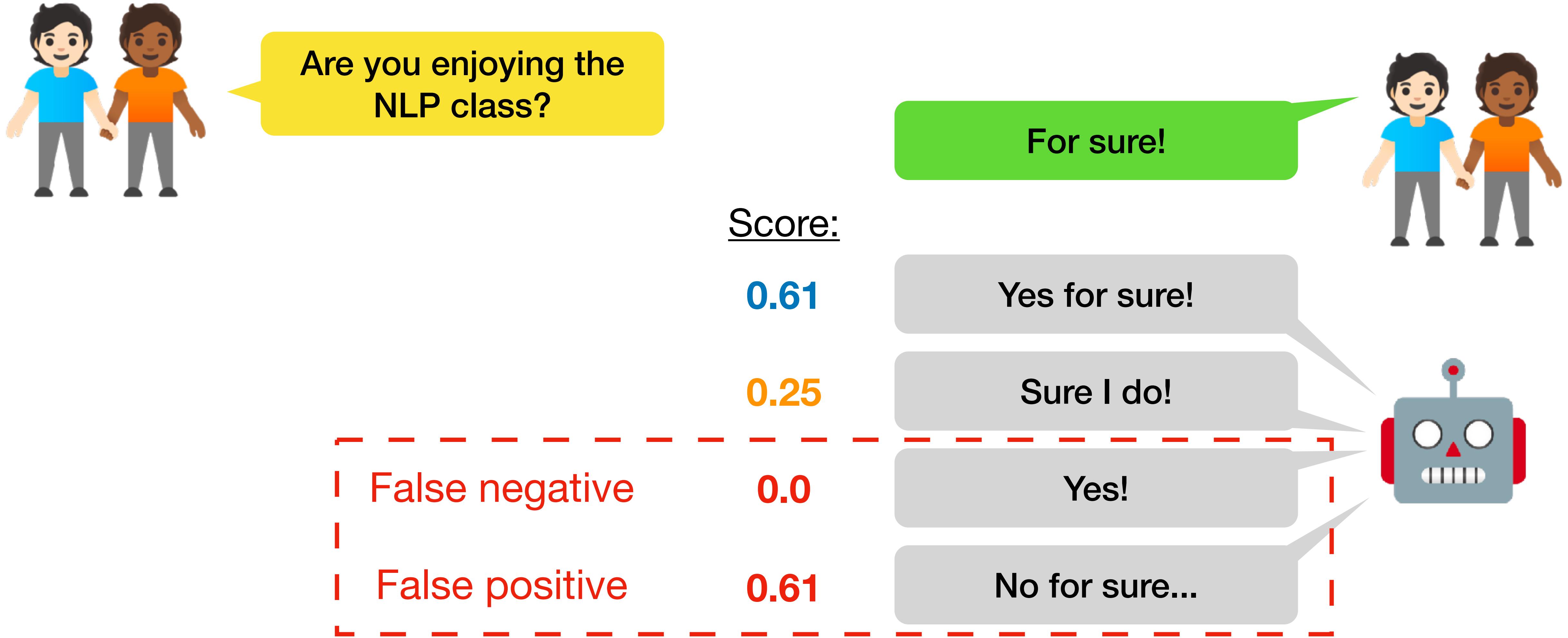
- Compute a score that indicates the similarity between *generated* and *gold-standard* (often human-written) text
- Fast and efficient; widely used (e.g. for MT and summarization)
- Dominant approach: *N-gram overlap* metrics
 - e.g., BLEU, ROUGE, METEOR, CIDEr, etc.

Content overlap metrics

- Dominant approach: *N-gram overlap* metrics
 - e.g., BLEU, ROUGE, METEOR, CIDEr, etc.
- Not ideal even for less open-ended tasks - e.g., machine translation
- They get progressively much worse for more open-ended tasks
 - Worse for summarization, as longer summaries are harder to measure
 - Much worse for dialogue (in how many ways can you respond to your friend?)
 - Much, much worse for story generation, which is also open-ended, but whose sequence length can make it seem you're getting decent scores!

A simple failure case

- N-gram overlap metrics have no concept of **semantic relatedness!**



A more comprehensive failure analysis

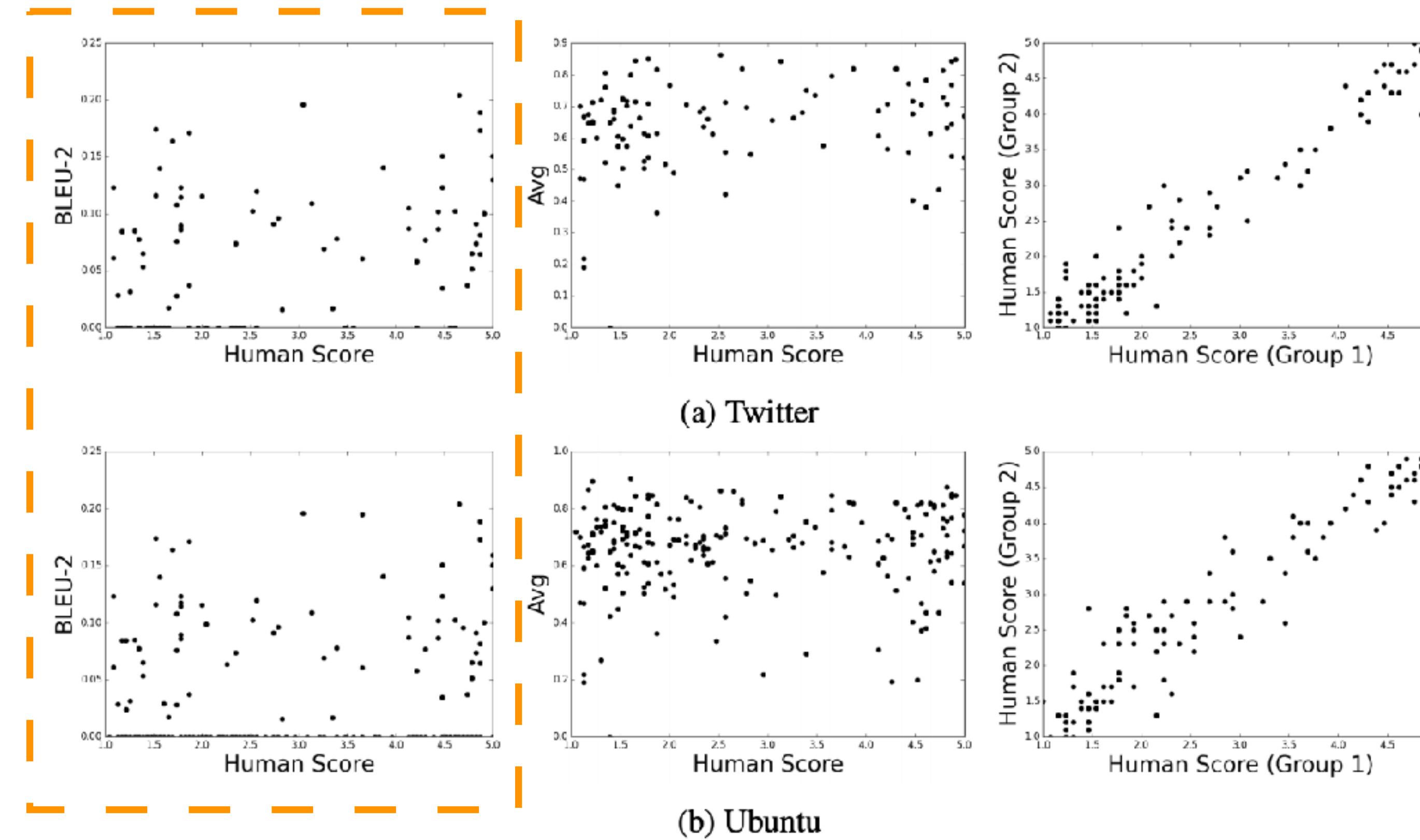
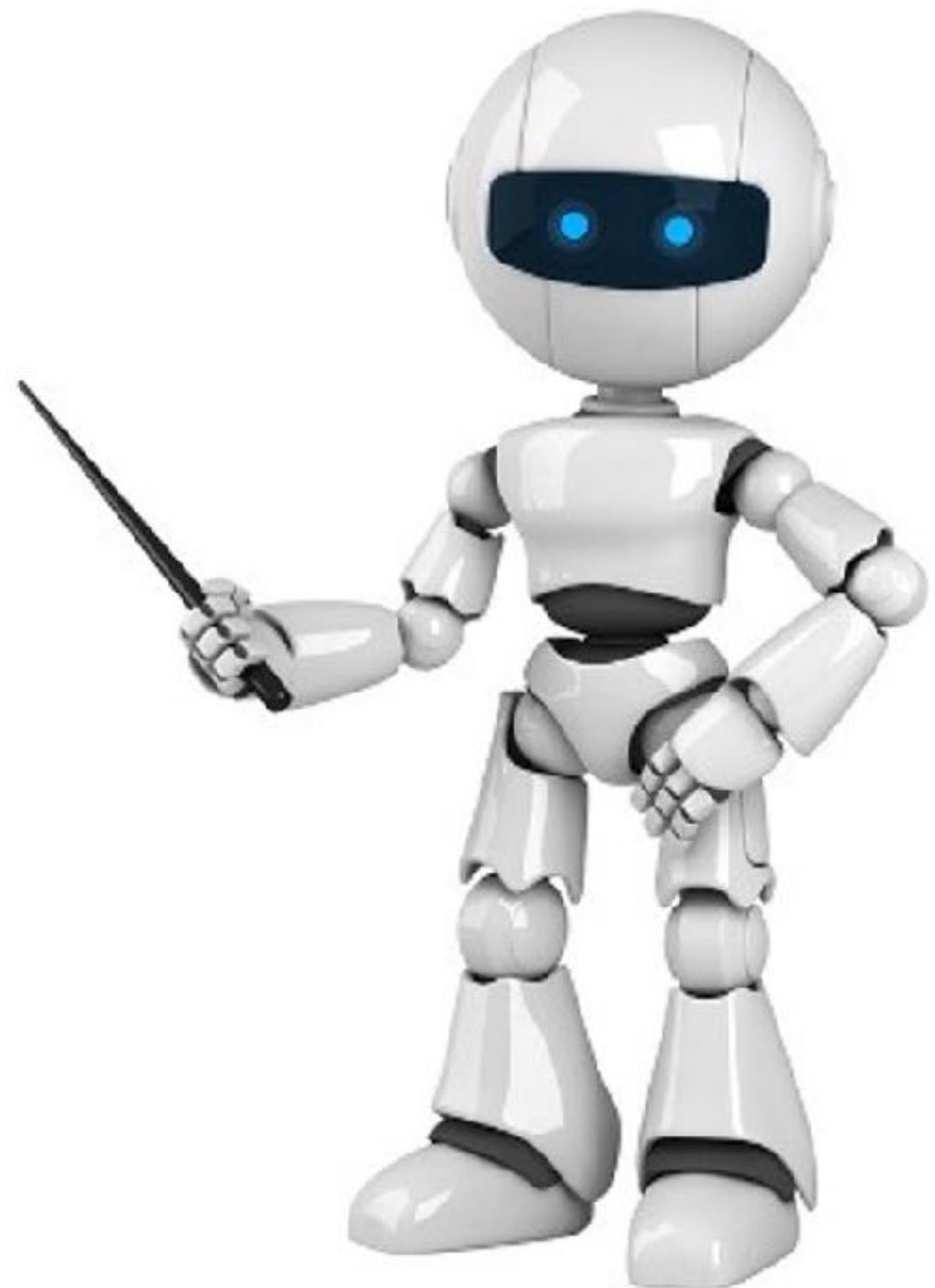


Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

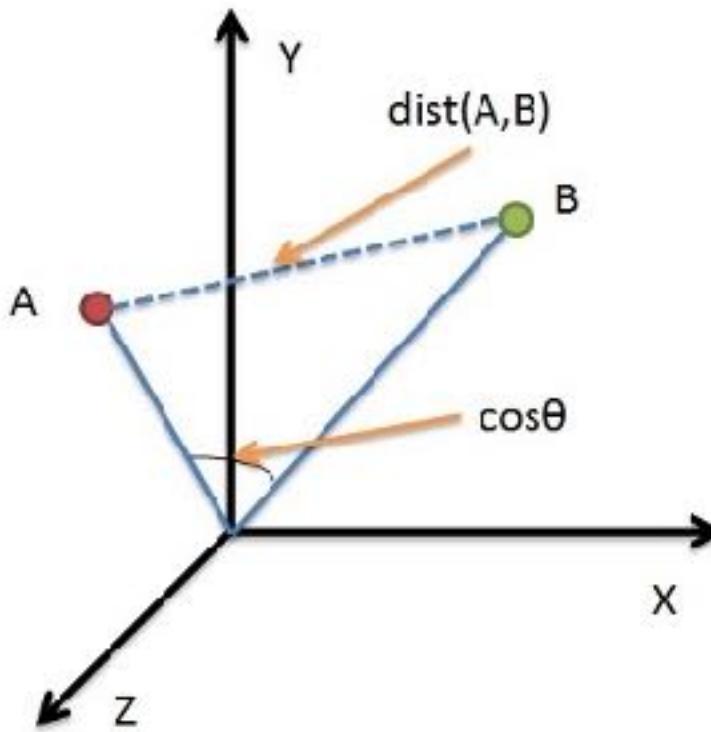
- Higher n -gram overlap does not imply higher **human score**.

Model-based metrics to capture more semantics

- Use learned representation of words and sentences to compute semantic similarity between generated and reference texts
- No more n-gram bottleneck: text units are represented as embeddings!
- Even though embeddings are pre-trained, distance metrics used to measure similarity can be fixed.



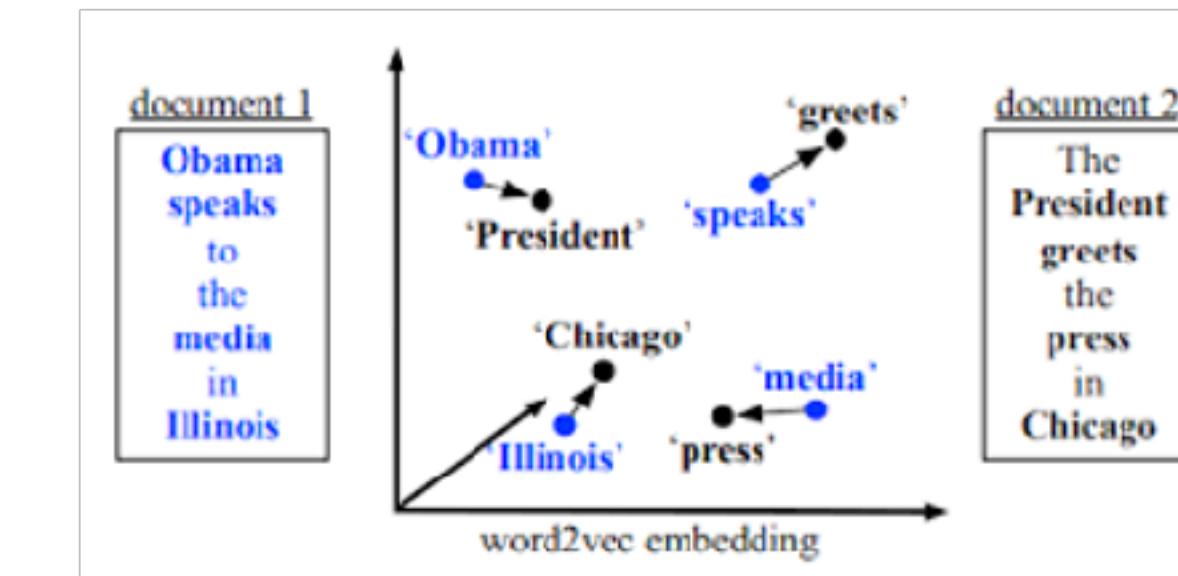
Model-based metrics: Word distance functions



Vector Similarity

Embedding-based similarity for semantic distance between text.

- Embedding Average (*Liu et al., 2016*)
- Vector Extrema (*Liu et al., 2016*)
- MEANT (*Lo, 2017*)
- YISI (*Lo, 2019*)



Word Mover's Distance

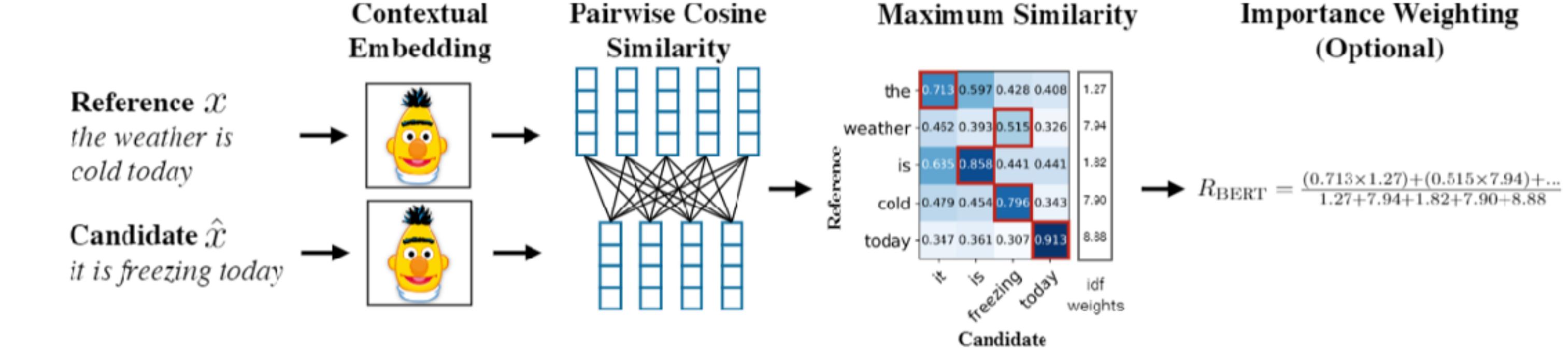
Measures the distance between two sequences using word embedding similarity matching.

- (*Kusner et al., 2015; Zhao et al., 2019*)

BERTSCORE

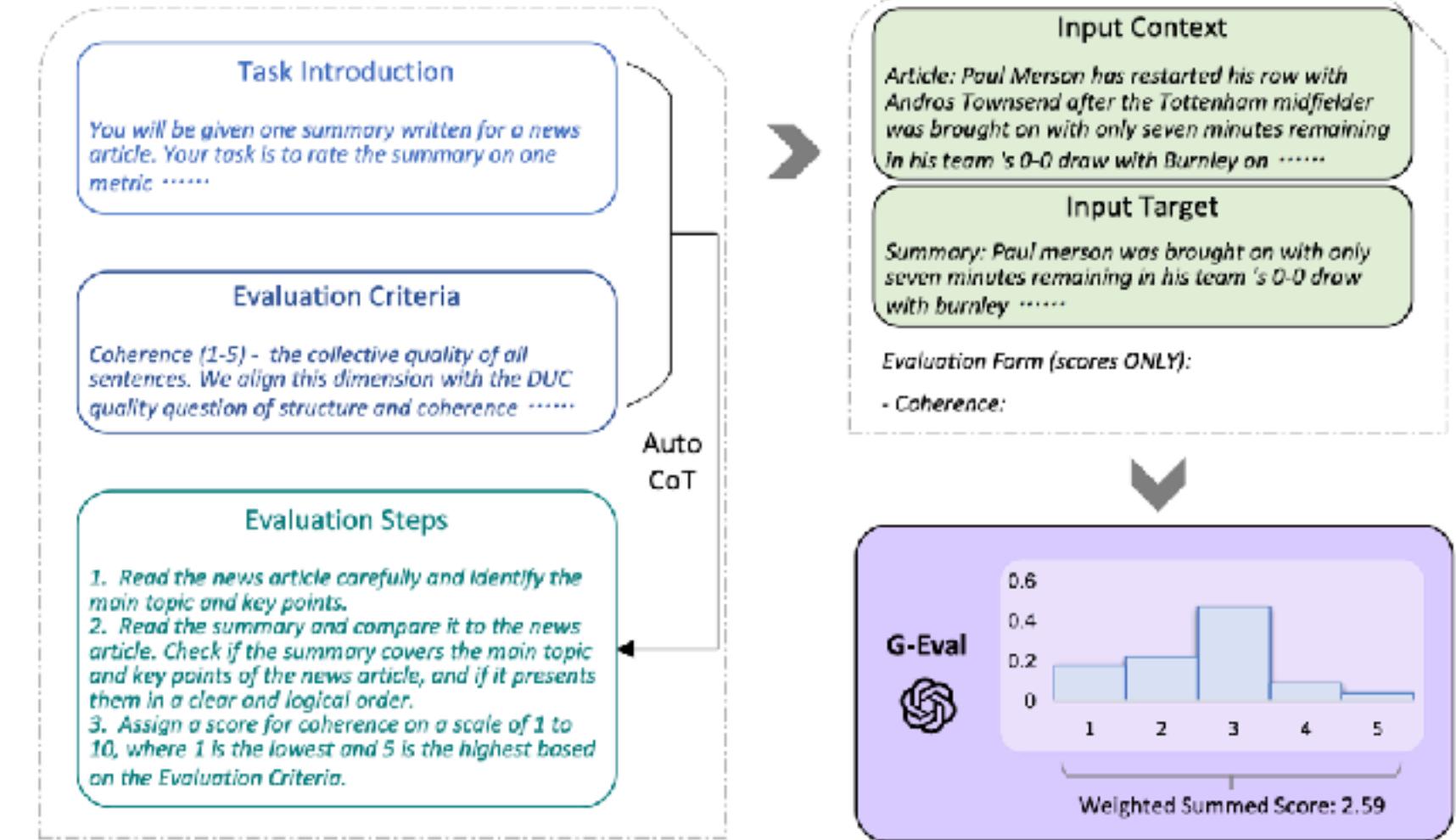
Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.

- (*Zhang et al., 2019*)



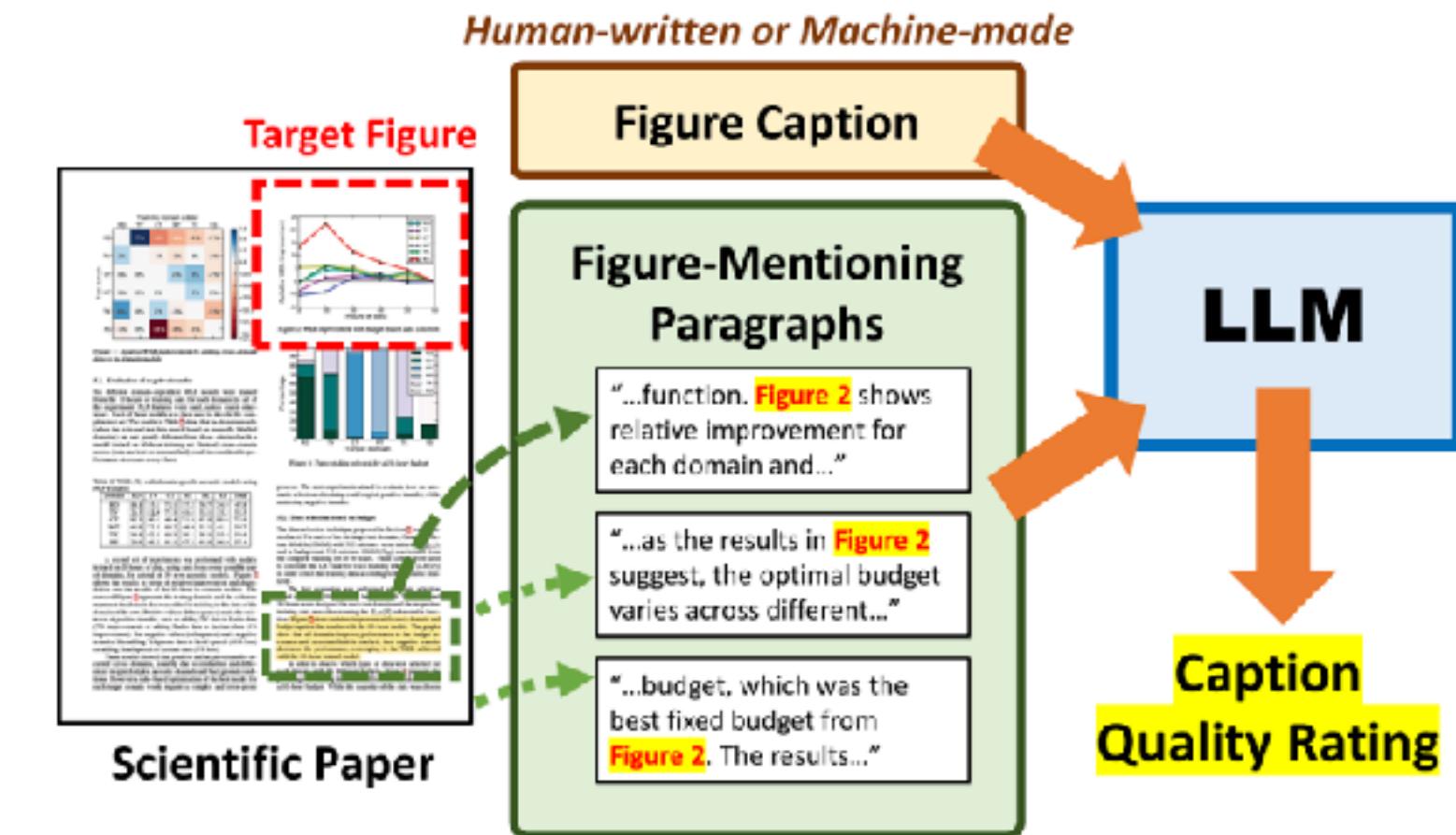
Model-based metrics: LLM as evaluator

- Directly prompt LLM (GPT-4) to evaluate generated text.
 - Can be **customized** with evaluation criteria
 - (Often) better correlation with human evaluators than task-specific metrics (e.g. ROUGE)
 - (Often) is **cheaper** than human evaluation



Liu et al. 2023

- Limitations
 - Brittleness: LLM evaluation can significantly vary when **given different prompts!**
 - Potential **self-bias** - LLMs may prefer what LLMs have generated...



Hsu et al. EMNLP Findings, 2023

Human evaluations



- Automatic metrics fall short of matching human decisions
- Most important form of evaluation for text generation systems
- Gold standard in developing new automatic metrics
 - Better automatic metrics will better correlate with human judgements!

Human evaluations

- Sounds easy, but hard in practice: [Ask humans](#) to evaluate the quality of text
- Typical evaluation dimensions:
 - fluency
 - coherence / consistency
 - factuality and correctness
 - commonsense
 - style / formality
 - grammaticality
 - typicality
 - redundancy
 - ...

Note: Don't compare human evaluation scores across different studies

Even if they claim to evaluate on the same dimensions!

Human evaluations

- Sounds easy, but hard in practice: [Ask humans](#) to evaluate the quality of text
- Typical evaluation dimensions:
 - fluency
 - coherence / consistency
 - factuality and correctness
 - commonsense
 - style / formality
 - grammaticality
 - typicality
 - redundancy
 - ...

Note: Don't compare human evaluation scores across different studies

Even if they claim to evaluate on the same dimensions!

Chatbot Arena

- Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference
- Try it out: <https://lmarena.ai>

LMSYS Chatbot Arena Leaderboard

[Vote](#) | [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over 400,000 human preference votes to rank LLMs with the Elo ranking system.

Total #models: 73. Total #votes: 408144. Last updated: March 13, 2024.

Contribute your vote [here](#)! Find more analysis in the [notebook](#).

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4.1106-prefixed	1251	+5/-4	48226	OpenAI	Proprietary	2023/4
1	GPT-4.0125-prefixed	1249	+5/-6	22282	OpenAI	Proprietary	2023/12
1	Claude-3.0Opus	1247	+6/-6	14854	Anthropic	Proprietary	2023/8
4	Rard_(Gemini_Pra)	1202	+6/-7	12623	Google	Proprietary	Online
4	Claude-3.Sonnet	1199	+6/-6	14845	Anthropic	Proprietary	2023/8
5	GPT-4.0314	1185	+4/-6	27245	OpenAI	Proprietary	2021/9
7	GPT-4.0413	1159	+4/-5	43763	OpenAI	Proprietary	2021/9
7	Mistral-Large-24B	1155	+5/-6	18959	Mistral	Proprietary	Unknown
8	Qwen1.5-72B-Chat	1146	+4/-5	16729	Alibaba	Qianwen LICENSE	2024/2
8	Claude-1	1145	+5/-6	21929	Anthropic	Proprietary	Unknown
8	Mistral-Medium	1145	+5/-4	23931	Mistral	Proprietary	Unknown

[Arena \(battle\)](#) [Arena \(side-by-side\)](#) [Direct Chat](#) [Vision Direct Chat](#) [Leaderboard](#) [About Us](#)

Chatbot Arena: Benchmarking LLMs in the Wild

[Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)

Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can continue chatting until you identify a winner.
- Vote won't be counted if model identity is revealed during conversation.

Arena Elo Leaderboard

We collect 400K+ human votes to compute an Elo-based LLM leaderboard. Find out who is the 🏆 LLM Champion!

Chat now!

Expand to see the descriptions of 32 models

Model A Model B

when was hku founded?

The University of Hong Kong (HKU) was founded in 1911. It is the oldest tertiary education institution in Hong Kong. HKU was established as a colonial-style university with the mission of providing higher education to the local Chinese population and promoting the development of Hong Kong. Today, HKU is a leading research-intensive university in Asia, known for its strong academic programs and research excellence in various fields.

when was hku founded?

Hong Kong University (HKU) was founded in 1911 and is the oldest tertiary institution in Hong Kong. It was initially established as the Hong Kong College of Medicine for Chinese, which was founded by the British medical missionary James Cantlie. Throughout its history, the university has undergone several name changes and expansions. In 1927, it was renamed the University of Hong Kong, reflecting its development into a full-fledged university offering various degree programs. HKU has played a significant role in the development of Hong Kong, contributing to its growth in education, research, and public service. It is highly respected and is known for its academic excellence, being consistently ranked among the top universities in Asia and the world.

Model A: mistral-8x7b-instruct-v0.1 Model B: command-r

A is better B is better Tie Both are bad

Enter your prompt and press ENTER

New Round Regenerate Share

Send

Evaluation: Takeaways

- *Content-overlap metrics* provide a good starting point for evaluating the generation quality, but they're **not good enough on their own**
- *Model-based metrics* can be more correlated with human judgment, but often are **not interpretable**
- Human judgments are critical
 - But humans are **inconsistent!**
- In many cases, the best judge of output quality is **YOU!**
 - **Look at the actual generations - don't just rely on numbers.**
 - **Publicly release large samples of outputs from your system!**

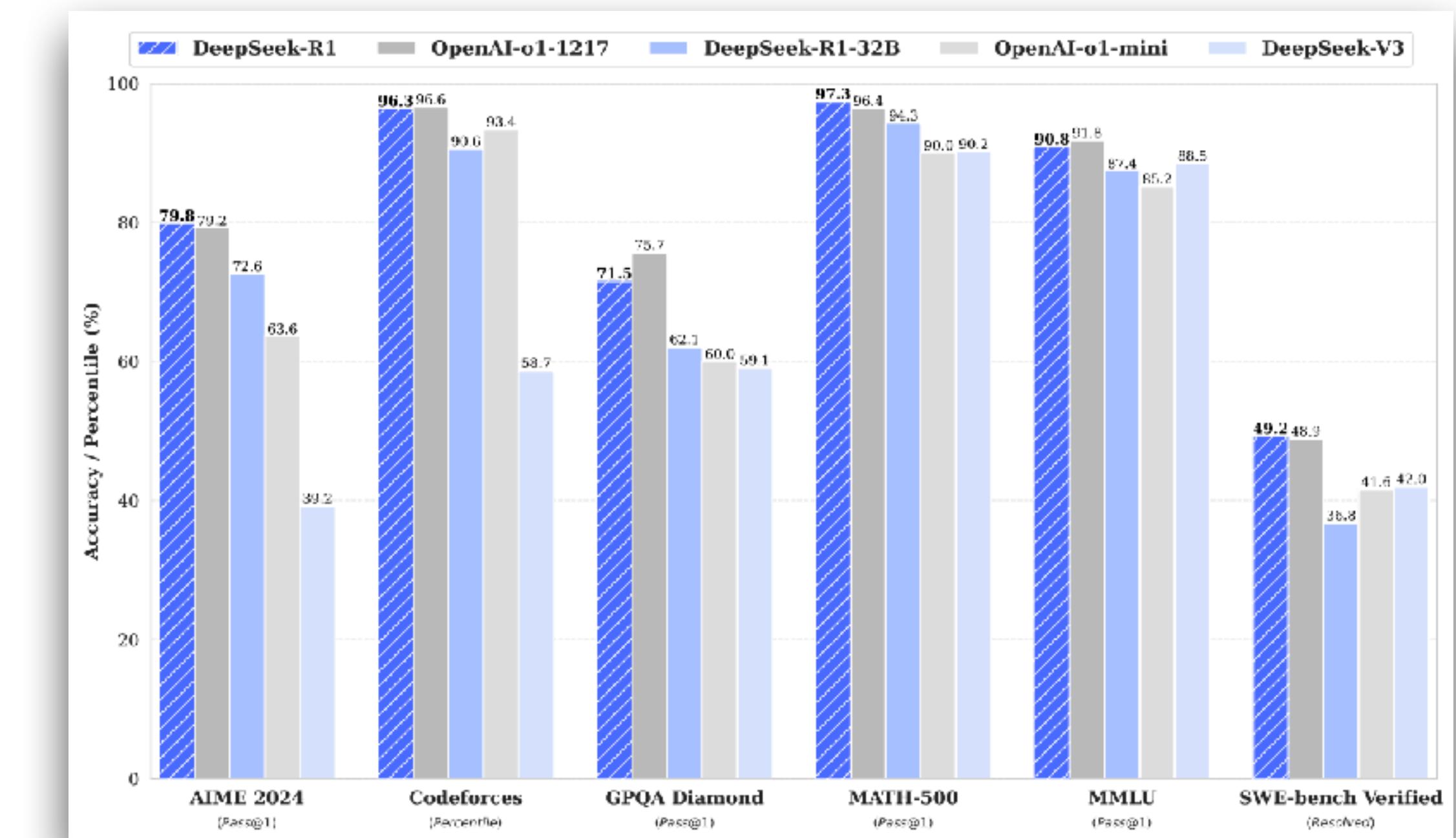
Concluding Thoughts

- Interacting with NLG systems quickly **shows their limitations**
- Even in tasks with more progress, there are **still many improvements ahead**
- Evaluation remains a huge challenge
 - We need better ways to **automatically evaluate NLG systems**
- One of the **most exciting areas** of NLP to work in!

LLM/VLM benchmarks and evaluation

LLM benchmarks

	Claude 3.7 Sonnet 64K extended thinking	Claude 3.7 Sonnet No extended thinking	Claude 3.5 Sonnet (new)	OpenAI o1 ¹	OpenAI o3-mini ¹ High	DeepSeek R1 32K extended thinking	Grok 3 Beta <i>Extended thinking</i>
Graduate-level reasoning GPQA Diamond ³	78.2% / 84.8%	68.0%	65.0%	75.7% / 78.0%	79.7%	71.5%	80.2% / 84.6%
Agentic coding SWE-bench Verified ²	—	62.3% / 70.3%	49.0%	48.9%	49.3%	49.2%	—
Agentic tool use TAU-bench	—	Retail 81.2%	Retail 71.5%	Retail 73.5%	—	—	—
—	—	Airline 58.4%	Airline 48.8%	Airline 54.2%	—	—	—
Multilingual Q&A MMMLU	86.1%	83.2%	82.1%	87.7%	79.5%	—	—
Visual reasoning MMMLU (validation)	75%	71.8%	70.4%	78.2 %	—	—	76.0% / 78.0%
Instruction-following IEEval	93.2%	90.8%	90.2%	—	—	83.3%	—
Math problem-solving MATH500	96.2%	82.2%	78.0%	96.4%	97.9%	97.3%	—
High school math competition AIME 2024 ⁴	61.3% / 80.0%	23.3%	16.0%	79.2% / 83.3%	87.3%	79.8%	83.9% / 93.3%



Multilingual QA

- MMMLU: Multilingual extension of Massive Multitask Language Understanding (MMLU)
 - Covering 57 subjects including elementary mathematics, US history, computer science, law, and more, with 15908 questions in total.

Conceptual Physics

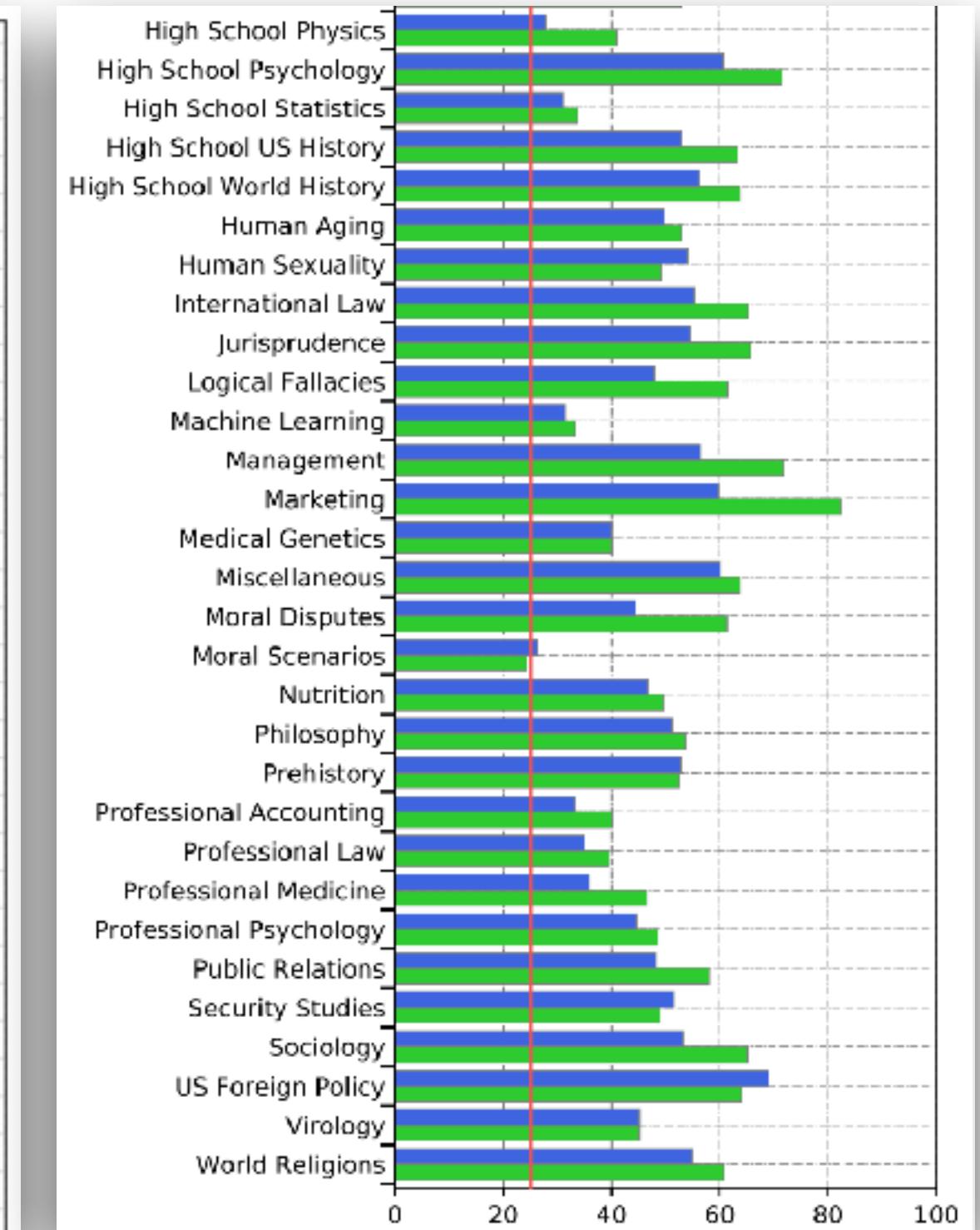
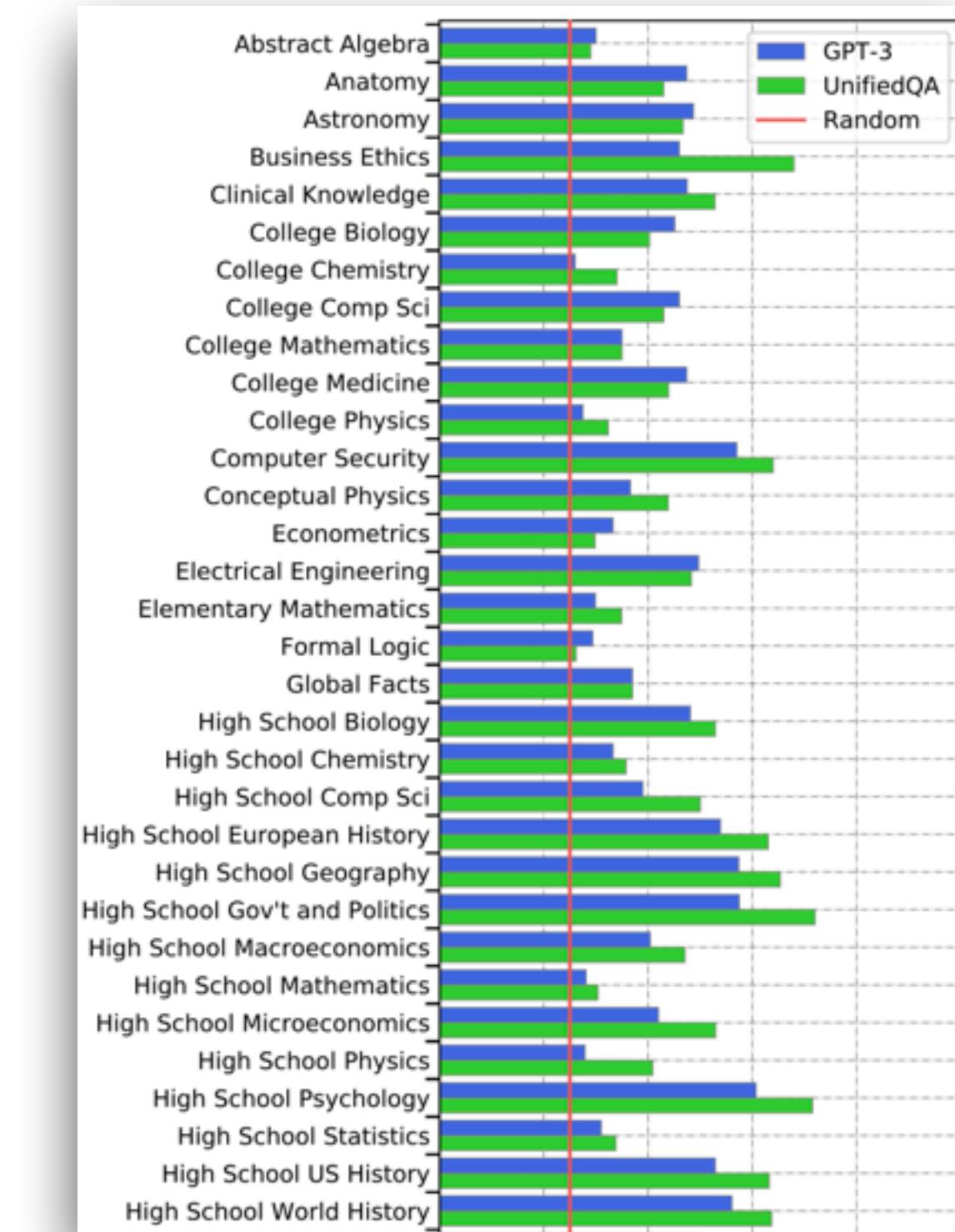
When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is

(A) 9.8 m/s^2
 (B) more than 9.8 m/s^2
 (C) less than 9.8 m/s^2
 (D) Cannot say unless the speed of throw is given.

College Mathematics

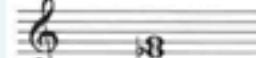
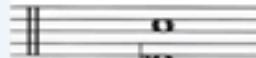
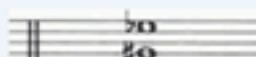
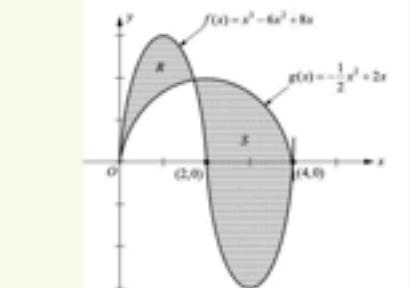
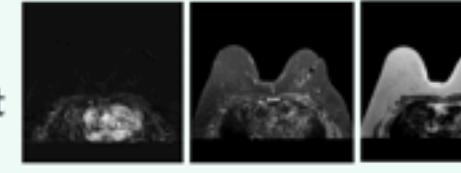
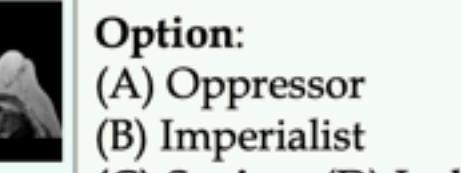
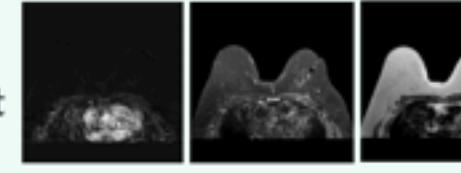
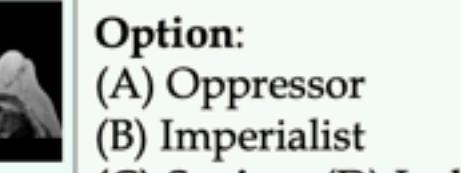
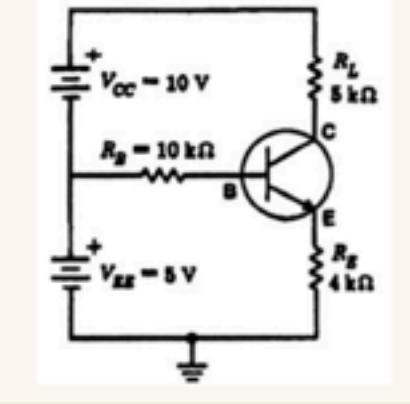
In the complex z -plane, the set of points satisfying the equation $z^2 = |z|^2$ is a

(A) pair of points
 (B) circle
 (C) half-line
 (D) line

Visual reasoning

- MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark
 - Including 11.5K meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines.

Art & Design	Business	Science
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <p>(A) Major third </p> <p>(B) Diminished fifth </p> <p>(C) Minor seventh </p> <p>(D) Diminished sixth </p>	<p>Question: ...The graph shown is compiled from data collected by Gallup . Find the probability that the selected Emotional Health Index Score is between 80.5 and 82?</p> <p>Options:</p> <p>(A) 0 (B) 0.2142</p> <p>(C) 0.3571 (D) 0.5</p>	<p>Question:  The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <p>(A) $\int_0^{1.5} [f(x) - g(x)] dx$</p> <p>(B) $\int_0^{1.5} [g(x) - f(x)] dx$</p> <p>(C) $\int_0^2 [f(x) - g(x)] dx$</p> <p>(D) $\int_0^2 [g(x) - x(x)] dx$</p>
<p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>
<p>Question: You are shown subtraction , T2 weighted  and T1 weighted axial  from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <p>(A) Susceptibility artifact </p> <p>(B) Hematoma </p> <p>(C) Fat necrosis </p> <p>(D) Silicone granuloma</p>	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? </p> <p>Option:</p> <p>(A) Oppressor</p> <p>(B) Imperialist</p> <p>(C) Savior</p> <p>(D) Isolationist</p>	<p>Question: Find the VCE for the circuit shown in .</p> <p>Answer: 3.75</p> <p>Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V</p>
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>

Math problem-solving

- MATH 500: 500 math problems- spanning topics like probability, algebra, trigonometry, and geometry.
- AIME 2024: Problems from the American Invitational Mathematics Examination (a prestigious high school mathematics competition) 2024

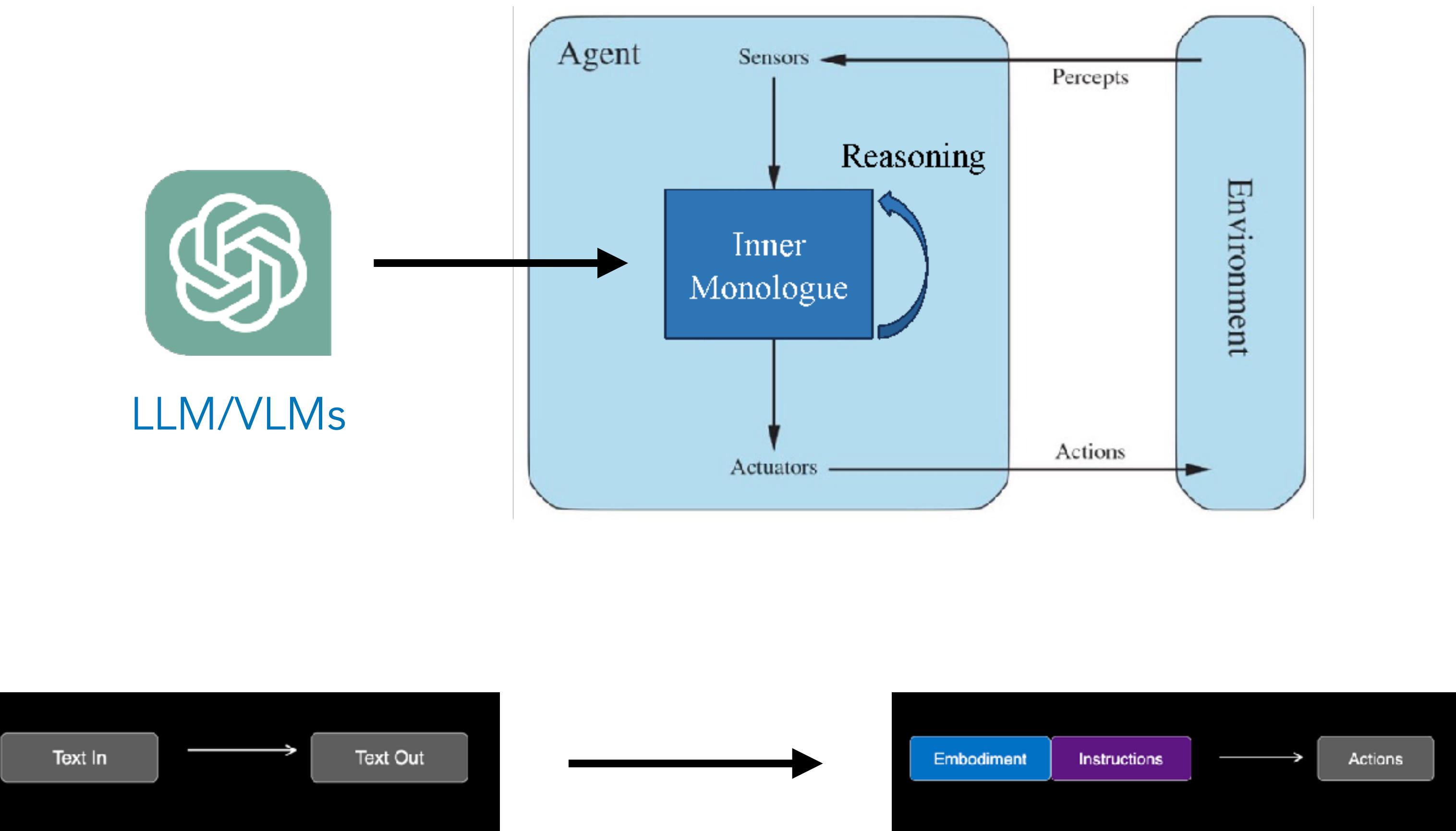
Problem	Solution	Answer	
string · lengths 333→405 23.3%	string · lengths 284→657 43.3%	int64 23→110 33.3%	<p>Convert the point $(0,3)$ in rectangular coordinates to polar coordinates. Enter your answer in the form (r,θ), where $r > 0$ and $0 \leq \theta < 2\pi$.</p> <p>We have that $r = \sqrt{0^2 + 3^2} = 3$. Also, if we draw the line connecting the origin and $(0,3)$, this line makes an angle of $\frac{\pi}{2}$ with the positive x-axis.</p> <pre>[asy] unitsize(0.8 cm); draw((-0.5,0)--(3.5,0)); draw((0,-0.5)--(0,3.5)); draw(arc((0,0),3,0,90),red,Arrow(6)); dot((0,3), red); label("(0,3)", (0,3), W); dot((3,0), red); [/asy]</pre> <p>Therefore, the polar coordinates are $\boxed{\left(3, \frac{\pi}{2} \right)}$.</p>
<p>Let x, y and z be positive real numbers that satisfy the following system of equations:</p> $\begin{aligned}\log_2\left(\frac{x}{yz}\right) &= \frac{1}{2} \\ \log_2\left(\frac{y}{xz}\right) &= \frac{1}{3} \\ \log_2\left(\frac{z}{xy}\right) &= \frac{1}{4}\end{aligned}$ <p>Then the value of $\log_2(x^4y^3z^2)$ is $\frac{m}{n}$ where m and n are relatively prime positive integers. Find $m+n$.</p>	<p>Denote $\log_2(x) = a$, $\log_2(y) = b$, and $\log_2(z) = c$.</p> <p>Then, we have:</p> $\begin{aligned}a-b-c &= \frac{1}{2}, \\ -a+b-c &= \frac{1}{3}, \\ -a-b+c &= \frac{1}{4}.\end{aligned}$ <p>Now, we can solve to get $a = \frac{-7}{24}$, $b = \frac{-9}{24}$, $c = \frac{-5}{12}$.</p> <p>Plugging these values in, we obtain $4a + 3b + 2c = \frac{25}{8} \implies \boxed{033}$.</p>	33	

Graduate-level reasoning

- GPOA
 - a challenging dataset of 448 multiple-choice questions written by domain experts in biology, physics, and chemistry.

Chemistry (general)
A reaction of a liquid organic compound, which molecules consist of carbon and hydrogen atoms, is performed at 80 centigrade and 20 bar for 24 hours. In the proton nuclear magnetic resonance spectrum, the signals with the highest chemical shift of the reactant are replaced by a signal of the product that is observed about three to four units downfield. Compounds from which position in the periodic system of the elements, which are also used in the corresponding large-scale industrial process, have been mostly likely initially added in small amounts? A) A metal compound from the fifth period. B) A metal compound from the fifth period and a non-metal compound from the third period. C) A metal compound from the fourth period. D) A metal compound from the fourth period and a non-metal compound from the second period.
Organic Chemistry
Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)? A) 2 B) 16 C) 8 D) 4

LLM/VLMs as agents

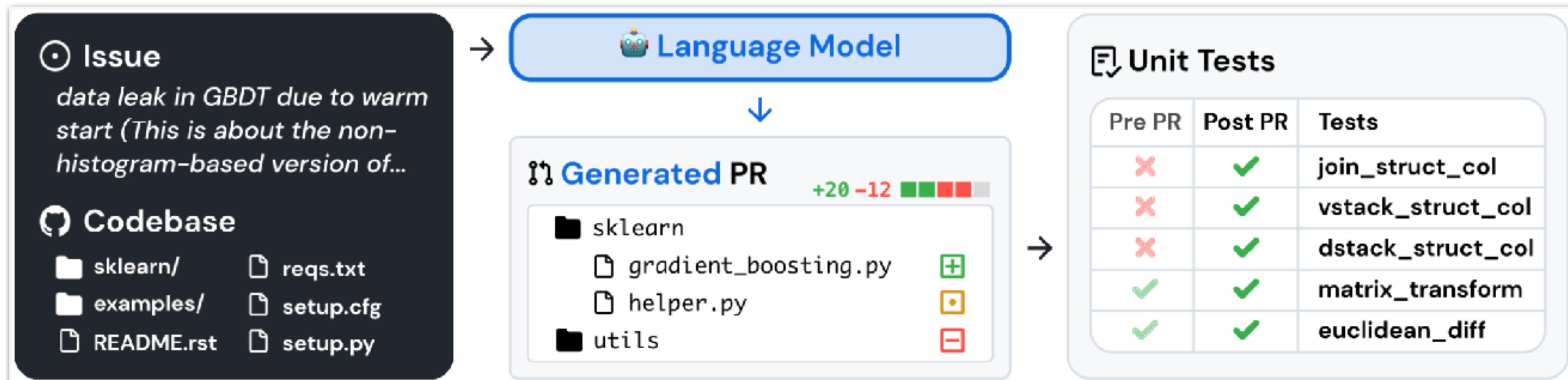


LLM/VLMs as agents

Benchmark type	Benchmark	Computer use (universal interface)		Web browsing agents	Human
		OpenAI CUA	Previous SOTA		
Computer use	OSWorld	38.1%	<u>22.0%</u>	-	<u>72.4%</u>
Browser use	WebArena	58.1%	<u>36.2%</u>	<u>57.1%</u>	<u>78.2%</u>
	WebVoyager	87.0%	56.0%	87.0%	-

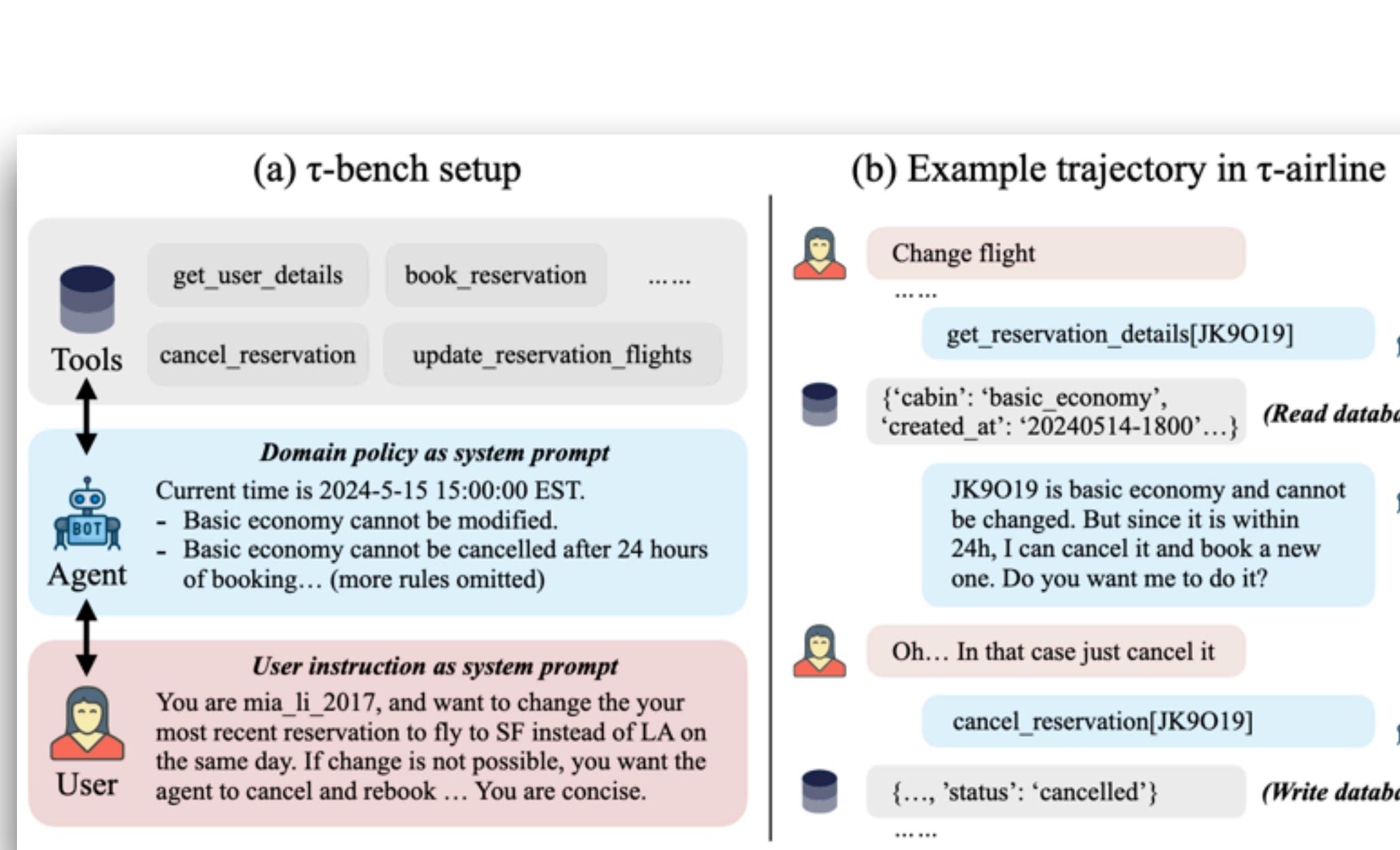
Agents coding

- SWE-bench: Project-level coding tasks
 - Environment: project code repos, filesystems, IDEs...
 - Observation space: code files, running outputs, docs, errors/issues, commit history...
 - Action space: code edits, file search/view, test updates...



Agentic tool use

- TAU-bench: API/functional calling for tool use
 - Environment: software systems such as databases, app/web services...
 - Observation space: API docs, system info, error messages and logs...
 - Action space: function calls, error handling routines...



(a) An orders database entry in τ-retail.

```
{
    "order_id": "#W2890441",
    "user_id": "mei_davis_8935",
    "items": [
        {
            "name": "Water Bottle",
            "product_id": "B310926033",
            "item_id": "2366567022",
            "price": 54.04,
            "options": [
                {
                    "capacity": "1000Oml",
                    "material": "stainless steel",
                    "color": "blue"
                }
            ],
            ...
        }
    ]
}
```

(b) An API tool in τ-retail.

```

def return_delivered_order_items(
    order_id: str,
    item_ids: List[str],
    payment_method_id: str,
) -> str: ...

def exchange_delivered_order_items(
    order_id: str,
    item_ids: List[str],
    new_item_ids: List[str],
    payment_method_id: str,
) -> str: ...

```

(c) Domain policy excerpts in τ-retail.

```

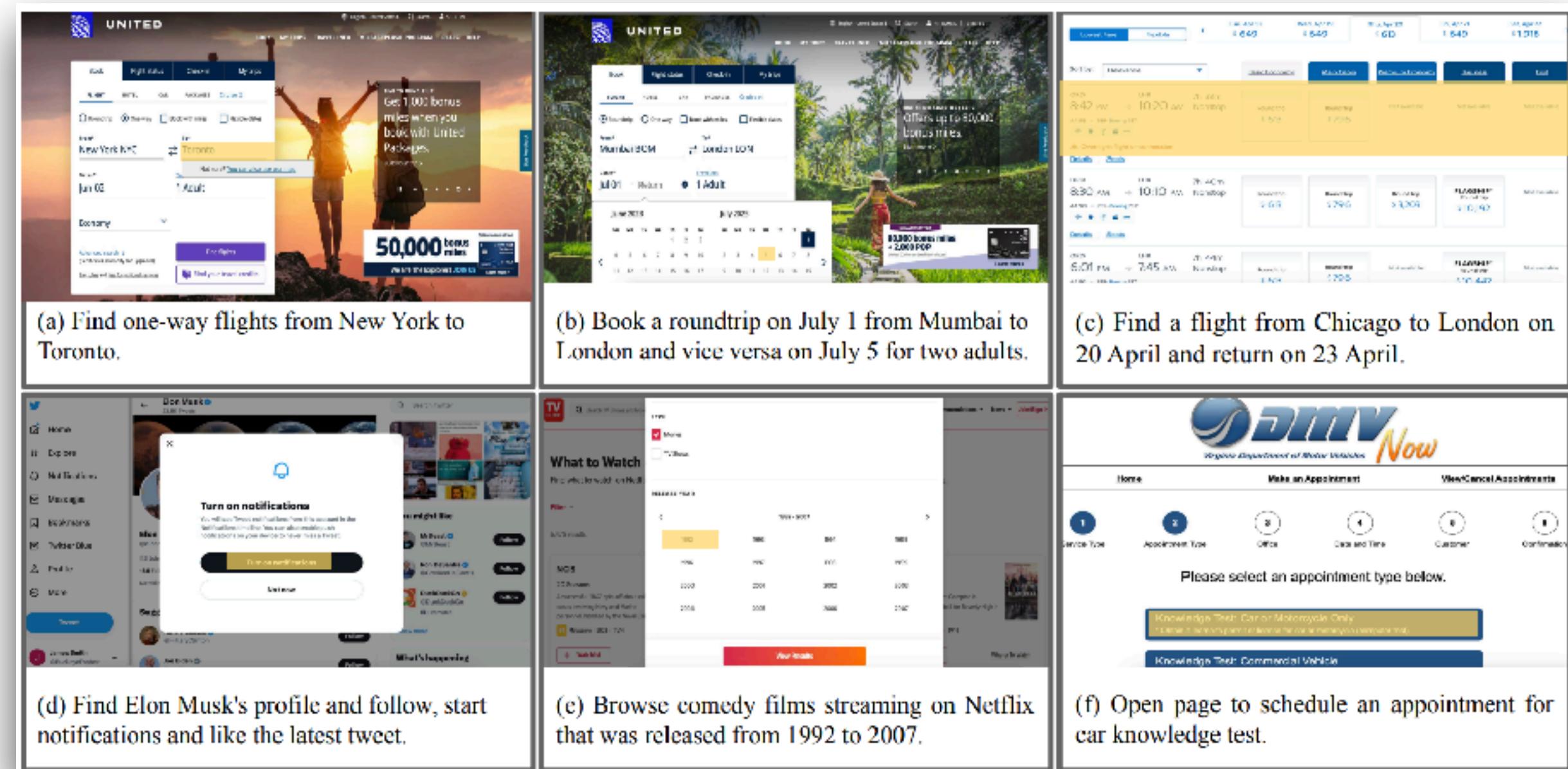
{
    "instruction": "You are Mei Davis in 80217. You want to return the water bottle, and exchange the pet bed and office chair to the cheapest version. Mention the two things together. If you can only do one of the two things, you prefer to do whatever saves you most money, but you want to know the money you can save in both ways. You are in debt and sad today, but very brief.",
    "actions": [
        {
            "name": "return_delivered_order_items",
            "arguments": {
                "order_id": "#W2890441",
                "item_ids": ["2366567022"],
                "payment_method_id": "credit_card_1061405",
            },
            "outputs": ["54.04", "41.64"]
        }
    ]
}

```

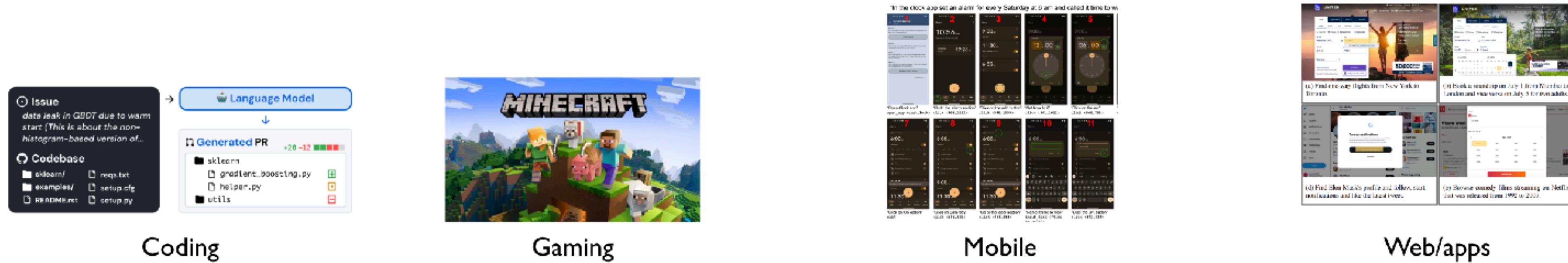
(d) User instruction ensures only one possible outcome.

Web/app agents

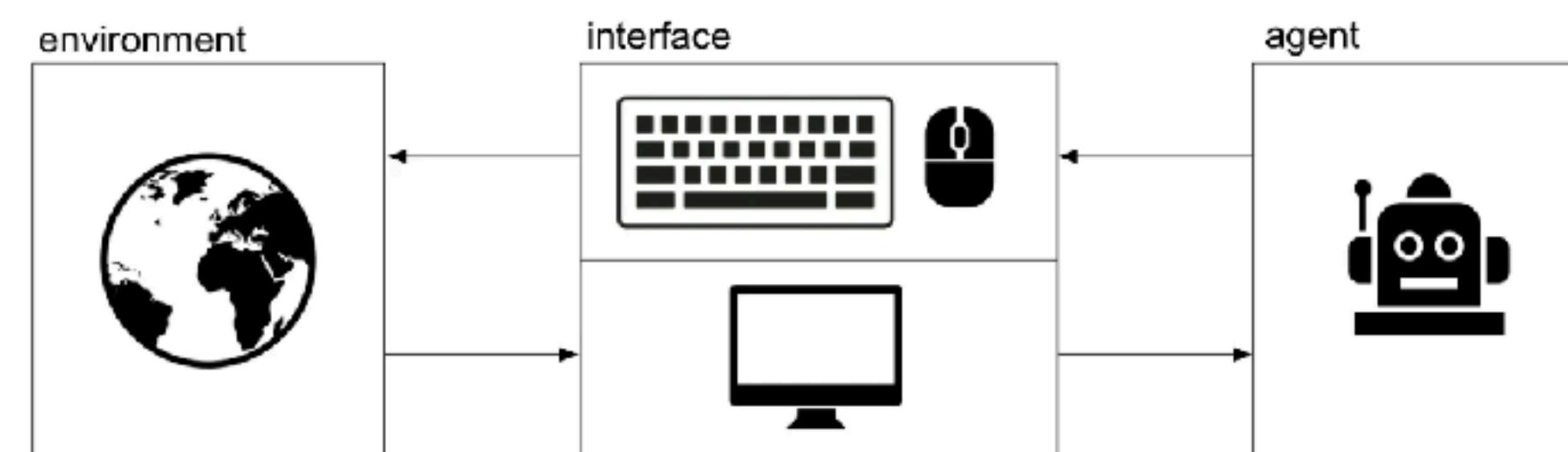
- WebArena: LLM/VLMs as agents for web/app use
 - Environment: web browsers/apps
 - Observation space: screenshots, DOM trees, HTML, historical actions...
 - Action space: browser/app controls (e.g., click, type, scroll, drag, hover...)



Universal digital environment?

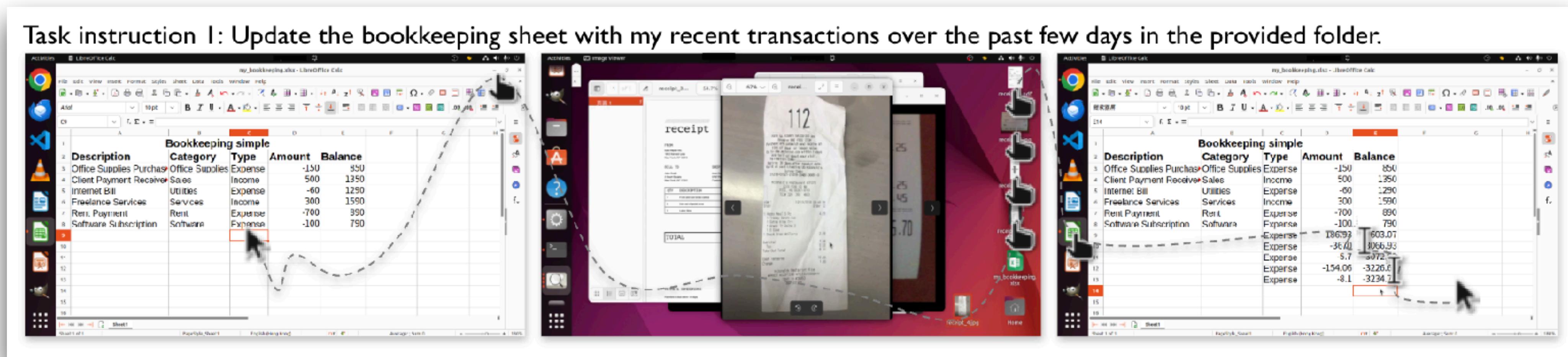


Can we study all digital AI agents in a **single** environment with **unified** observation and action spaces?



Computer use agents

- OSWorld: Computer use for universal digital tasks
 - Environment: desktop operating systems
 - Observation space: desktop screenshots, a11y trees, historical actions...
 - Action space: keyboard/mouse controls (e.g., click, type, drag, shortcuts)



Try it out: <https://arena.xlang.ai>

Agents in physical world: robotics

- Robotics for physical interaction
 - Environment: physical world spaces
 - Observation space: visual input, sensor readings, physical states, proprioception...
 - Action space: motor controls (e.g., move, grasp, manipulate...)Environment: desktop operating systems

