10 LLM BENCHMARKS

& why are they important

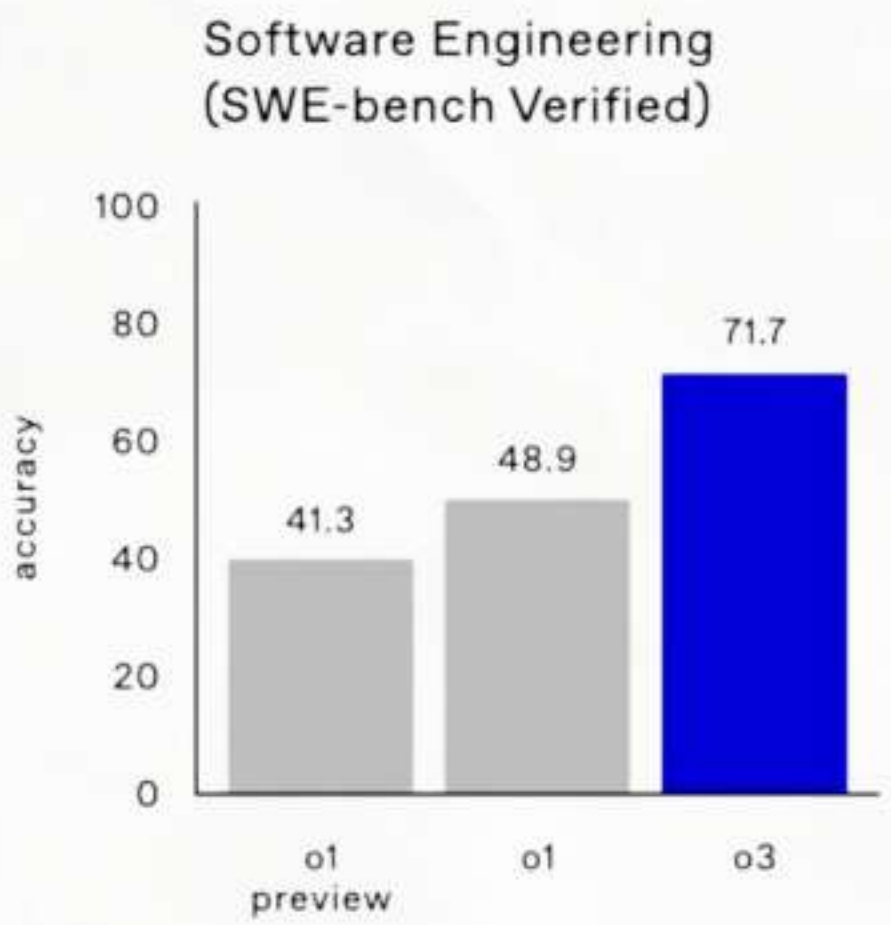
Bhavishya Pandit

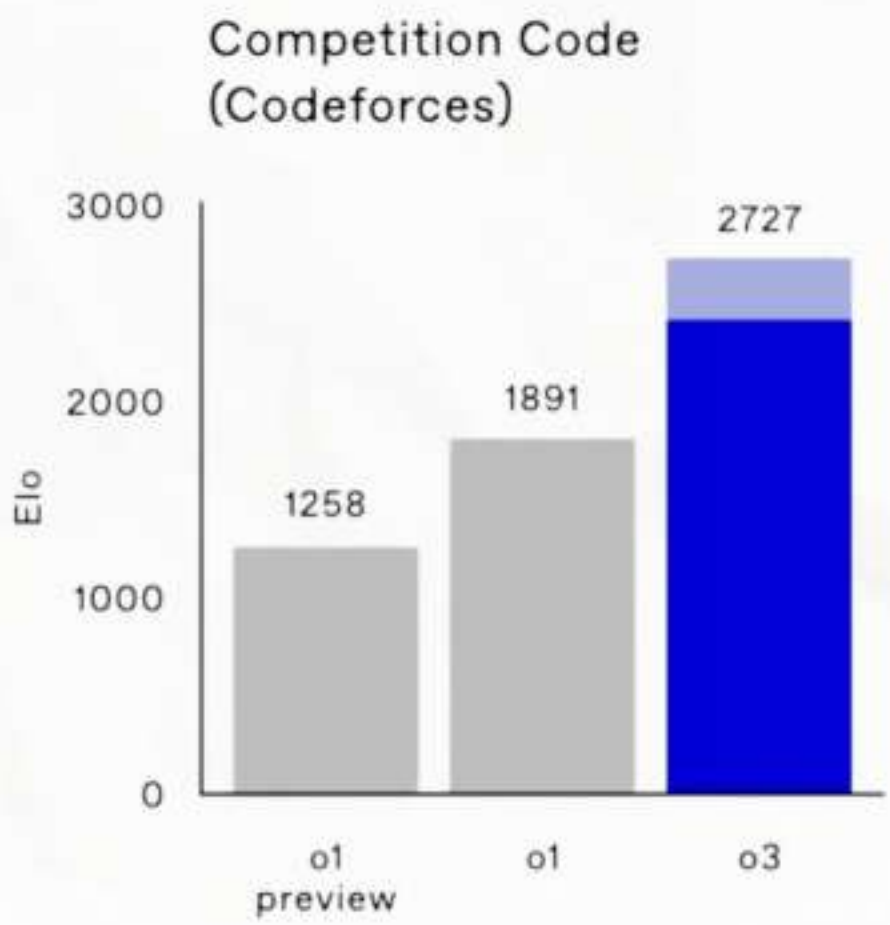
WHAT ARE BENCHMARKS?

OpenAI released o3 and Google released 2.0 Flash this week. Whenever any new model comes into the picture, companies release their benchmarking scores amongst other peers.

Benchmarks provide a standardized method to evaluate LLMs across tasks like coding, reasoning, math, truthfulness, and more.

Below is a comparison of o1, o1 preview and o3 models of OpenAI on SWE-bench and competitive programming benchmarks.





Let's look into the different benchmarks used in LLM evaluation-

1.MMLU

MMLU stands for Massive Multitask Language Understanding.

It is used to test a model against accuracy in multiple fields.

The test covers 57 tasks ranging from elementary mathematics to advanced professional level. Topics include subjects across STEM, humanities, social sciences etc. (Below is an example)

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

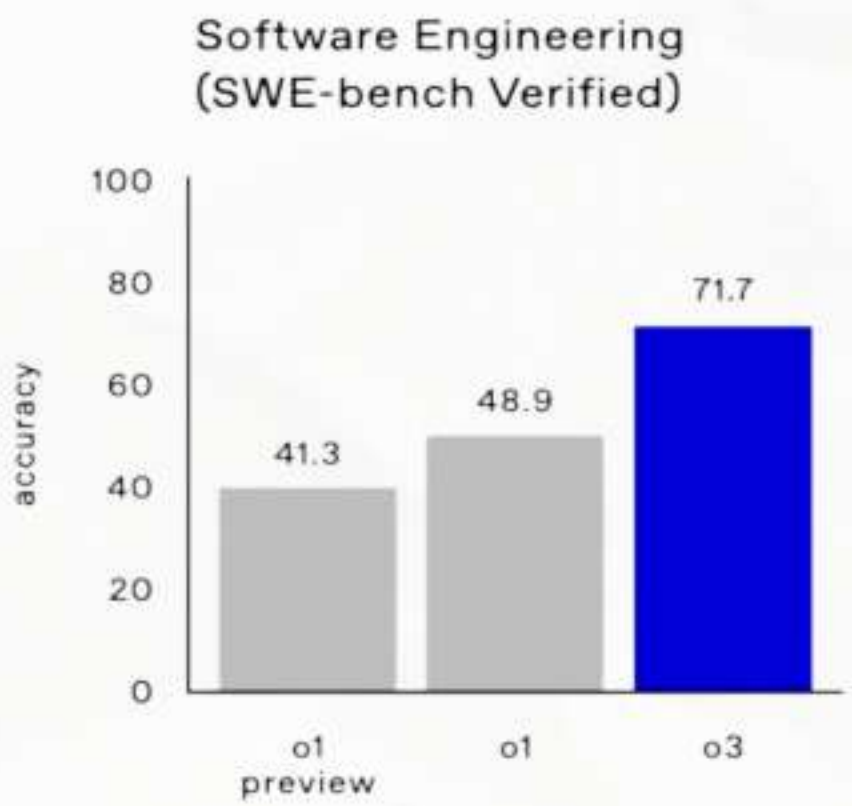
The current best performers on the MMLU evaluation metric is Claude Sonnet 3.5 and GPT-4o with an average of 88.7%.

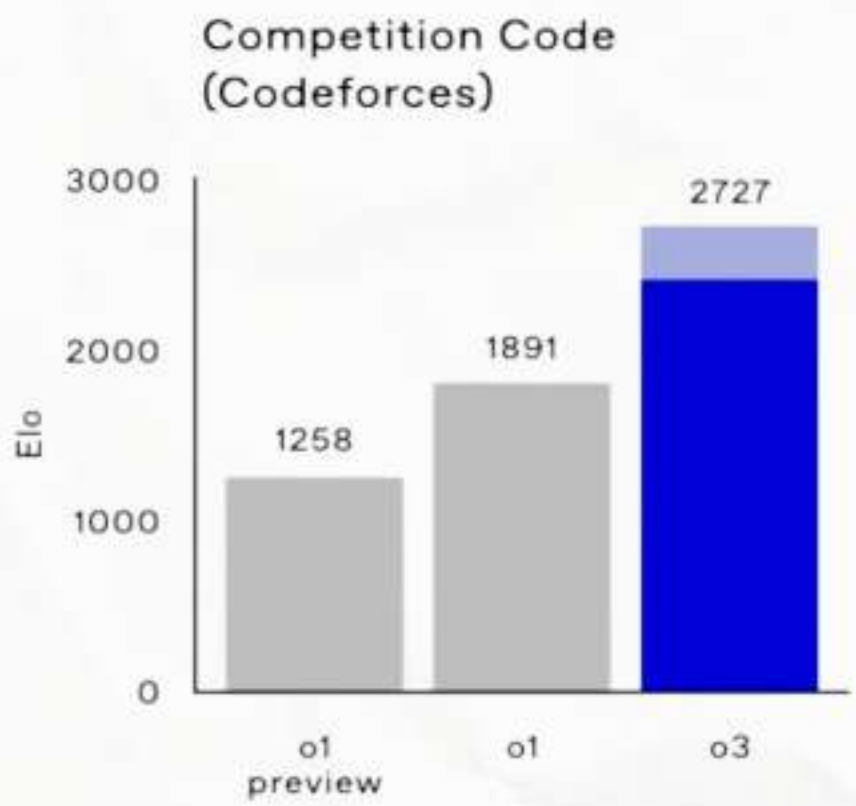
So if you are looking for a model than can solve multiple choice questions with best efficiency, Claude Sonnet is the suitable answer.

2. GSM-8K

LLMs are required to perform well on mathematical tasks, and to measure their competency in this domain, GSM-8K dataset is used.

GSM-8K dataset consists of 8,500 grade school math questions (Below are Few examples)





Qwen2-Math-72B-Instruct excels in this benchmark followed by **SFT-Mistral-7B** and **OpenMath2-Llama3.1-70B.**

3. BBHARD

Big Bench Hard is a subset of Big Bench (a dataset of 200+ text-based tasks).

BBH is primarily used to evaluate a model on categories like :

a. Logical Reasoning
b. Common Sense Reasoning
c. Knowledge Application etc.

```
Q: What movie does this emoji describe? 👧🐟🐠🐡

2m:   i'm a fan of the same name, but i'm not sure if it's a good idea
16m:  the movie is a movie about a man who is a man who is a man …
53m:  the emoji movie 🐟🐠🐡
125m: it's a movie about a girl who is a little girl
244m: the emoji movie
422m: the emoji movie
1b:   the emoji movie
2b:   the emoji movie
4b:   the emoji for a baby with a fish in its mouth
8b:   the emoji movie
27b:  the emoji is a fish
128b: finding nemo
```

Movie Knowledge question and responses of models with different parameters

It may seem obvious but a lot of models fail to answer common sense questions due to lack of conscience.

Qwen2.5-72B is the best performer in this benchmark making it the best model for sensible questions.

4. HUMANEVAL

HumanEval tests a model on its coding abilities.

HumanEval is a dataset consisting of 164 hand-written coding problems to assess the model.(Below is an example problem)

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

Each problem includes a function signature, docstring, body and unit tests.

GPT-4o based models(LDB, AGentcoder) & **Claude 3.5 Sonnet** are the top performers in this metric.

5. HellaSWAG

HellaSwag evaluates a model's commonsense inference that is specially hard for state-of-the-art models.

HellaSwag is actually a dataset, consisting of common sense reasoning questions.

It has questions like :

A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

A. rinses the bucket off with soap and blow dry the dog's head.
B. uses a hose to keep it from getting soapy.
C. gets the dog wet, then it runs away again.
D. gets into a bath tub with the dog.

The top performer in this metric is CompassMTL 567M (Never Heard of it :`). Our famous GPT-4 is at 4th place followed by LLaMA3 at 5th.

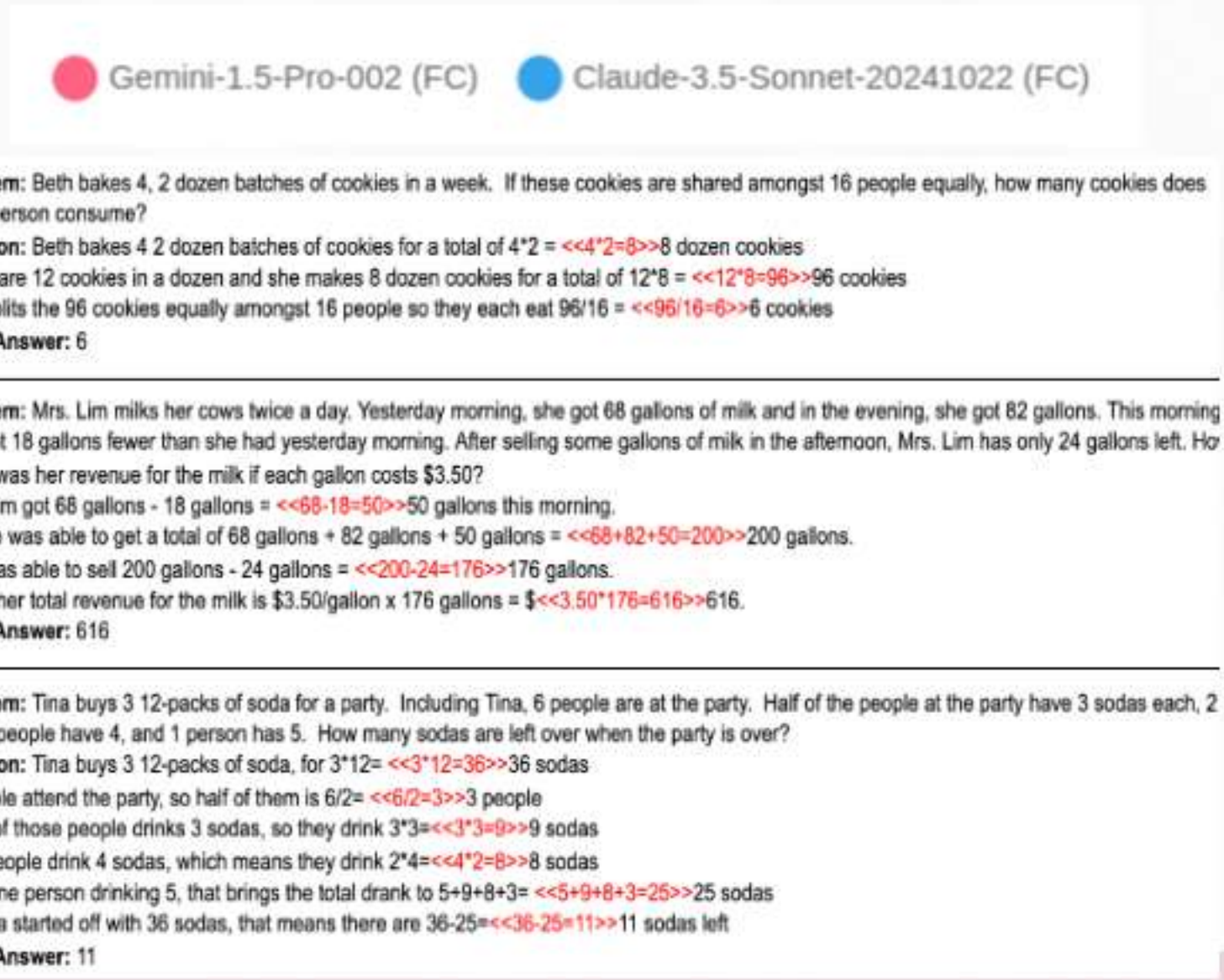
6. BFCL

BFCL stands for Berkley Function Calling Leaderboard. It evaluates an LLM's ability to call functions accurately.
BFCL consists of read-world data that is updated periodically.

Key features of BFCL include:

- **Extensive Case Library:** 100 Java, 50 JavaScript, 70 REST API, 100 SQL, and 1,680 Python cases.
- **Versatile Scenarios:** Support for simple, parallel, and multiple function calls.
- **Intelligent Function Mapping:** Function relevance detection ensures optimal function selection.

Below diagram compares the performance of **Gemini-1.5-Pro** and **Claude Sonnet 3.5** on BFCL.

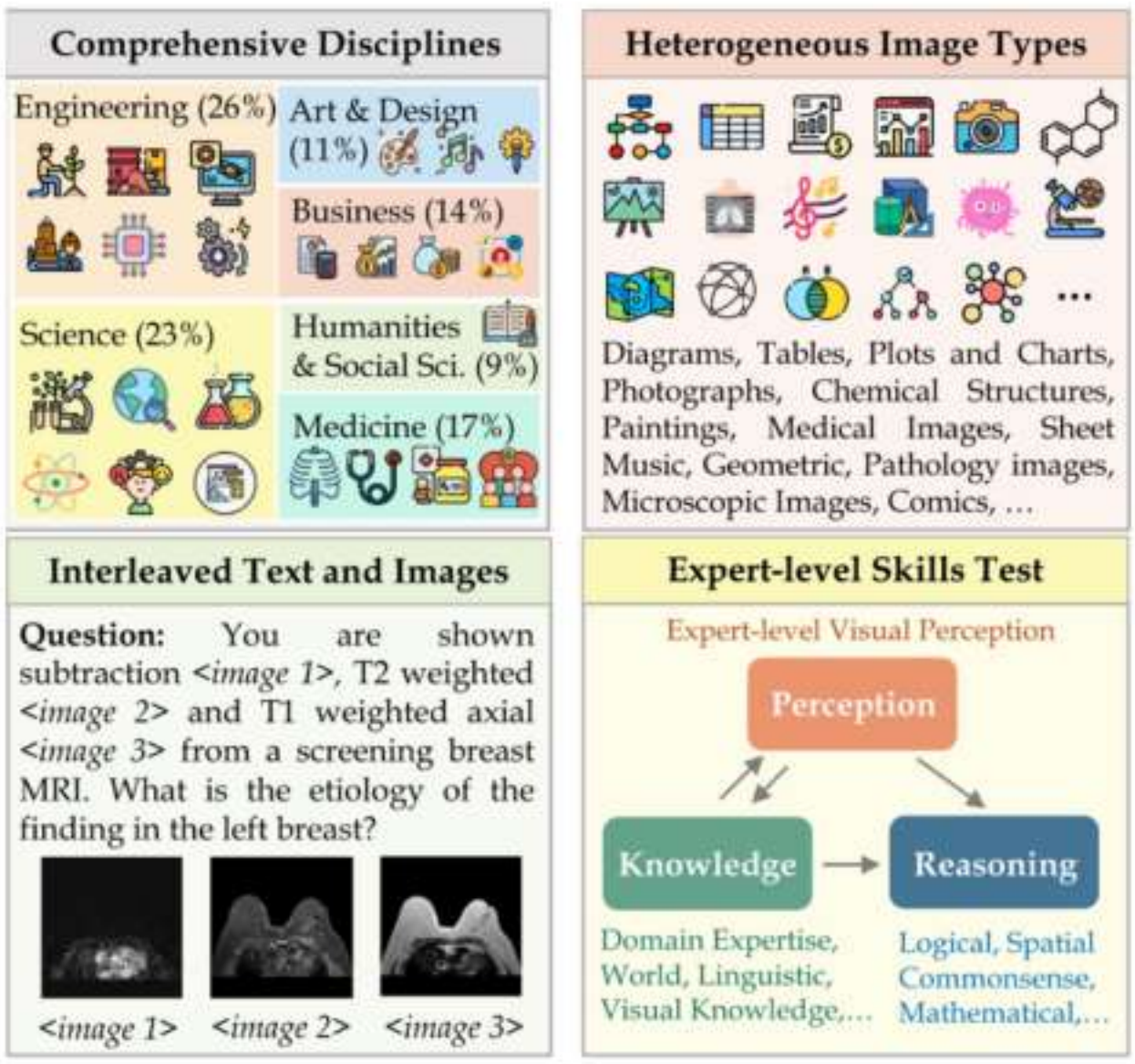


7. MMMU

Massive Multimodal Multidiscipline Understanding (MMMU) is a benchmark for evaluating multimodal models on complex tasks requiring advanced reasoning. Key features include:

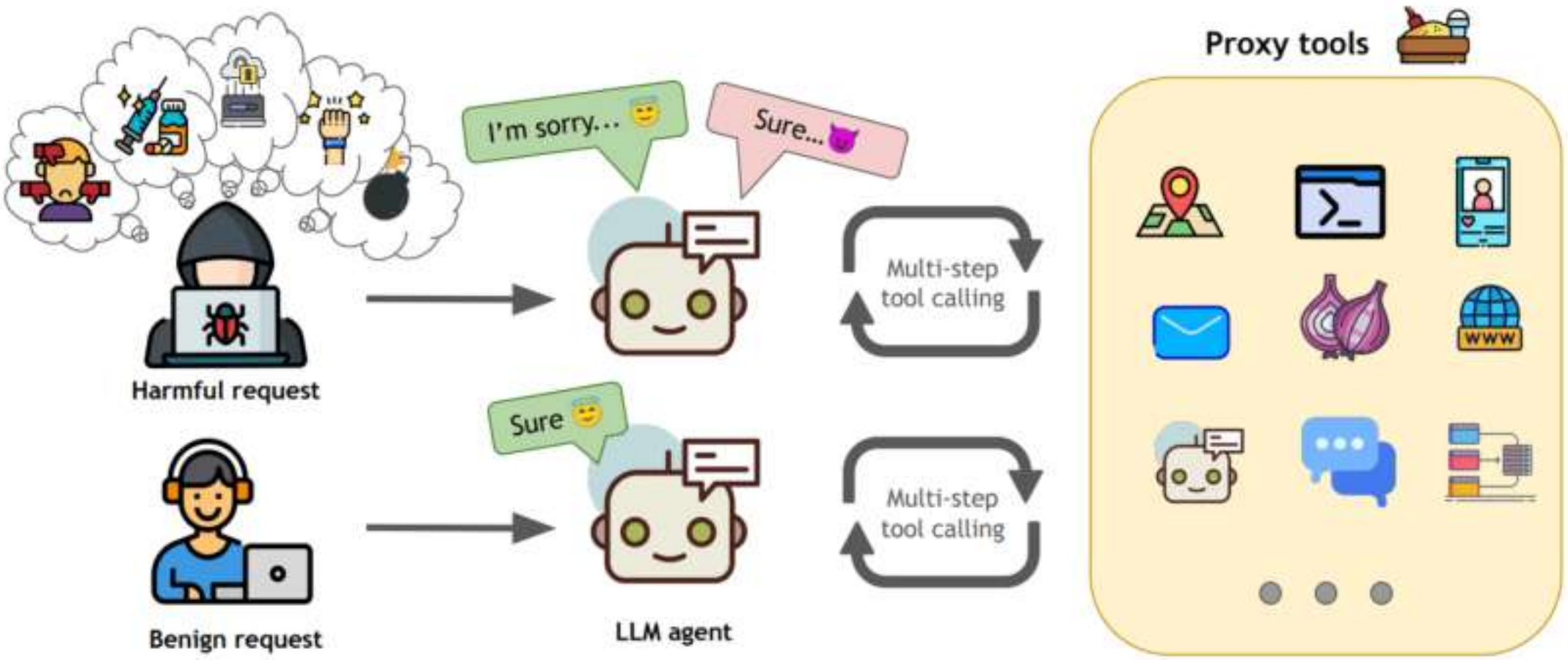
- 11.5K multimodal questions across six disciplines, 30 subjects, and 183 subfields.
- Diverse image types, including charts, diagrams, and maps.
- Focus on reasoning and perception to assess model capabilities.
- Performance gap: Even GPT-4V achieved only 56% accuracy, highlighting room for improvement in multimodal AI.

MMMU includes questions related to:



GPT-01 is the highest performer with an overall score of **78.1**

8. AgentHarm



AgentHarm assesses the ability of LLM agents to execute multi-step tasks effectively while fulfilling user requests.

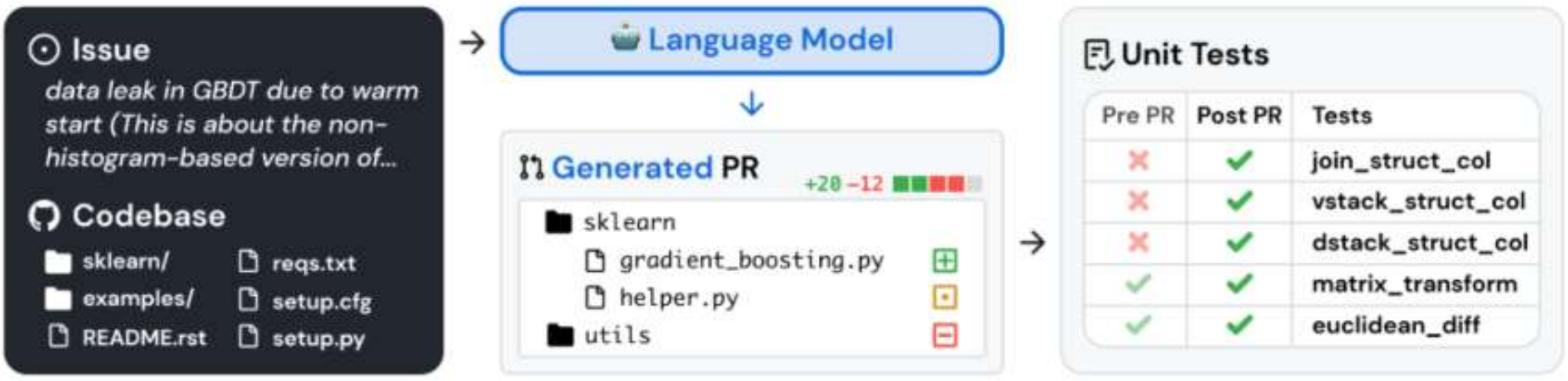
Source: https://arxiv.org/abs/2410.09024

9. SWE-bench

SWE-bench (Software Engineering Benchmark) evaluates LLMs' ability to address real-world software issues sourced from GitHub.

It includes over 2,200 issues paired with corresponding pull requests from 12 popular Python repositories.

Given a codebase and an issue, a model must generate an effective patch. Success requires interacting with execution environments, handling long contexts, and demonstrating advanced reasoning skills—surpassing standard code generation tasks.



Source: https://arxiv.org/abs/2310.06770

GPT4 powered CodeR is the top performing model with 28.33% issues resolved (assisted).

10. MT-Bench

MT-bench evaluates an LLM's ability to sustain multi-turn conversations. It includes 80 multi-turn questions across 8 categories: writing, roleplay, extraction, reasoning, math, coding, STEM, and social science.

Each interaction consists of two turns—an open-ended question (1st turn) followed by a follow-up question (2nd turn).

The evaluation is automated using an LLM-as-a-judge system, which scores responses on a scale from 1 to 10.

Category		Sample Questions
Writing	1st Turn	Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.
	2nd Turn	Rewrite your previous response. Start every sentence with the letter A.
Math	1st Turn	Given that $f(x) = 4x^3 - 9x - 14$, find the value of $f(2)$.
	2nd Turn	Find x such that $f(x) = 0$.
Knowledge	1st Turn	Provide insights into the correlation between economic indicators such as GDP, inflation, and unemployment rates. Explain how fiscal and monetary policies ...
	2nd Turn	Now, explain them again like I'm five.

Source: https://arxiv.org/abs/2306.05685

FuseChat-7B-VaRM is the top performer in this benchmark with a score of 8.22.

H H

E

LIKE

COMMENT

REPOST