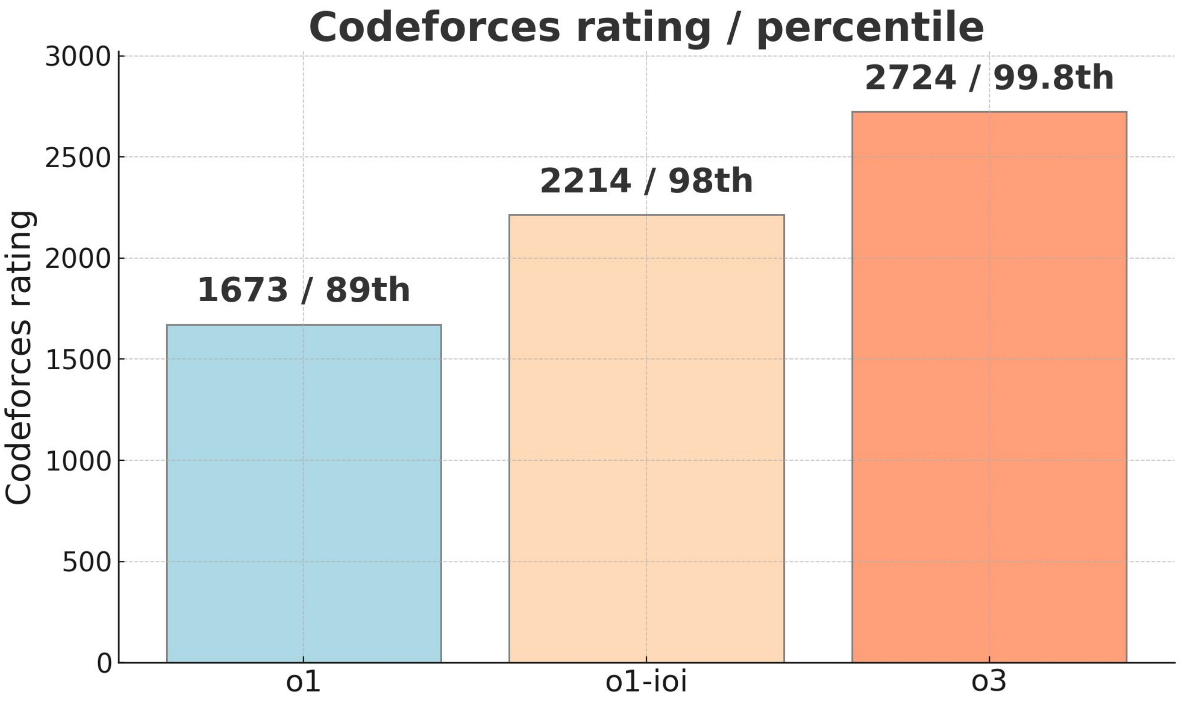
MLE-bench:

Evaluating Machine Learning Agents on Machine Learning Engineering

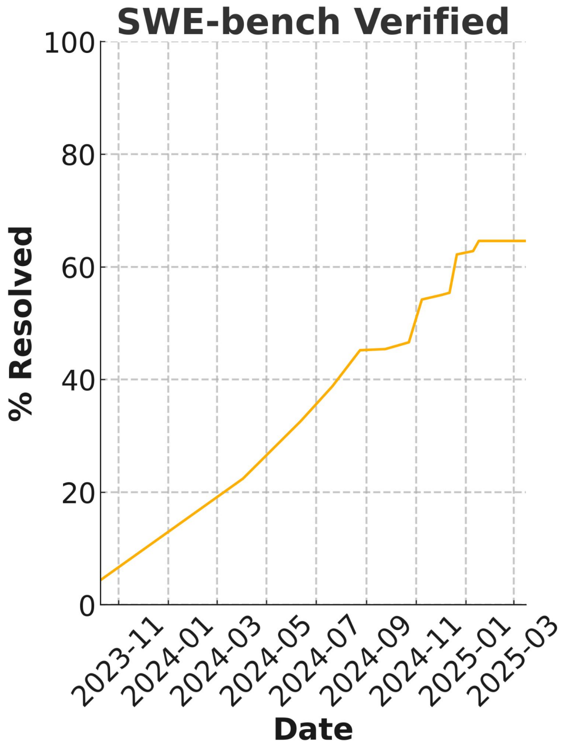
Chan Jun Shern*, Neil Chowdhury*, Oliver Jaffe*, James Aung*, Dane Sherburn*, Evan Mays*, Giulio Starace*,
Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, Aleksander Mądry

*Equal Contribution. Authors randomized.

LMs show impressive coding benchmark performance

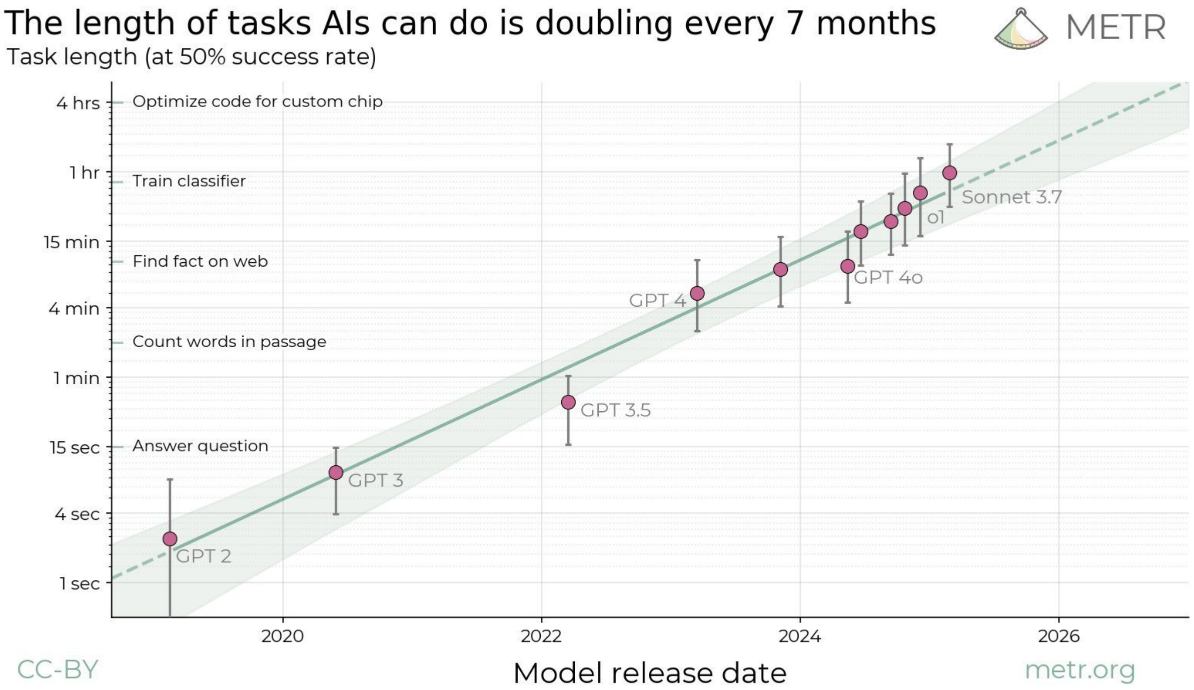


Source: Competitive Programming with Large Reasoning Models



Source: SWE-bench

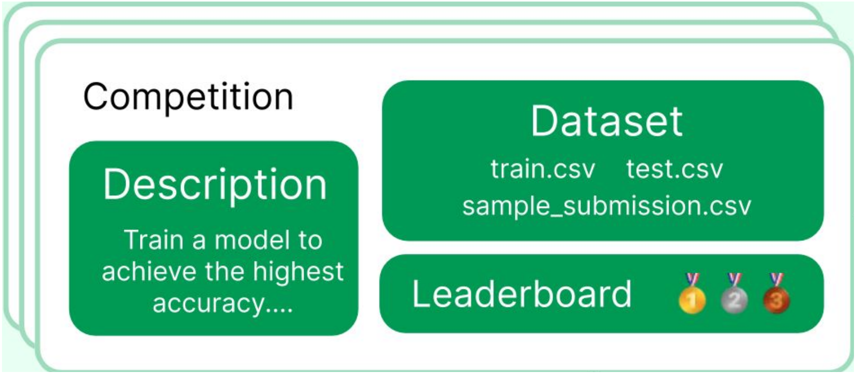
LMs show impressive coding benchmark performance



Source: Measuring AI Ability to Complete Long Tasks

But how good are ML agents at **ML engineering**?

MLE-bench: Evaluate on 75 Kaggle competitions



Competition

Description

Train a model to achieve the highest accuracy....

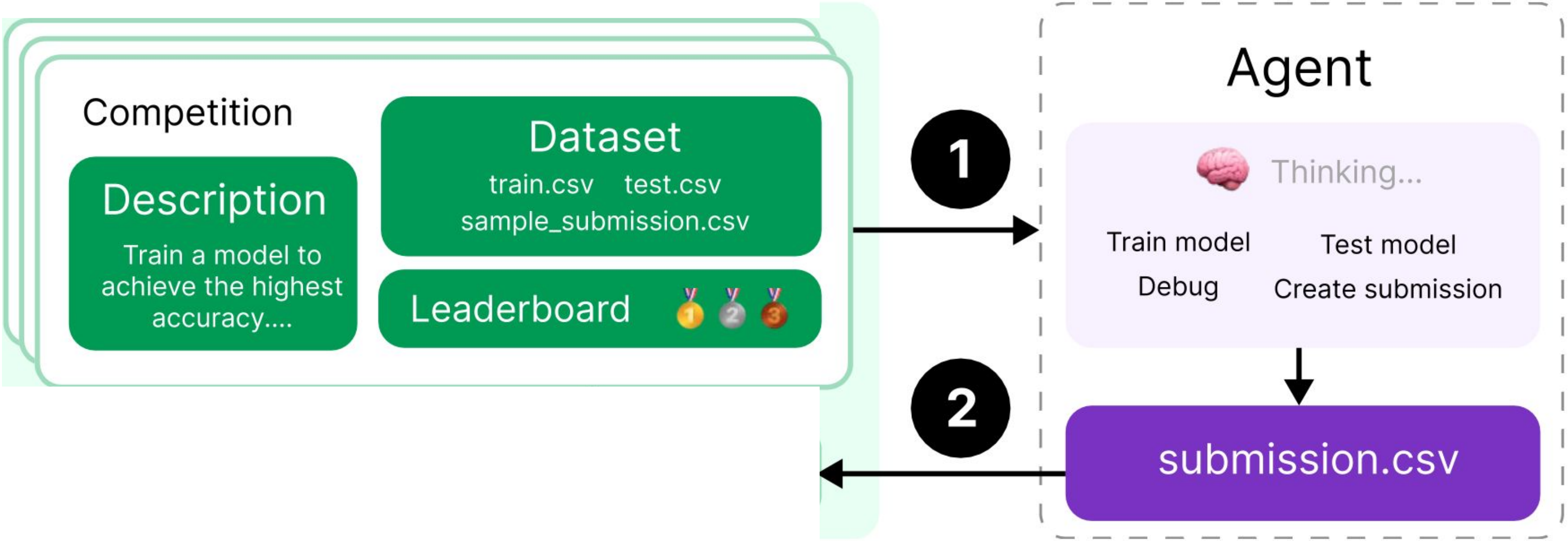
Dataset

train.csv test.csv

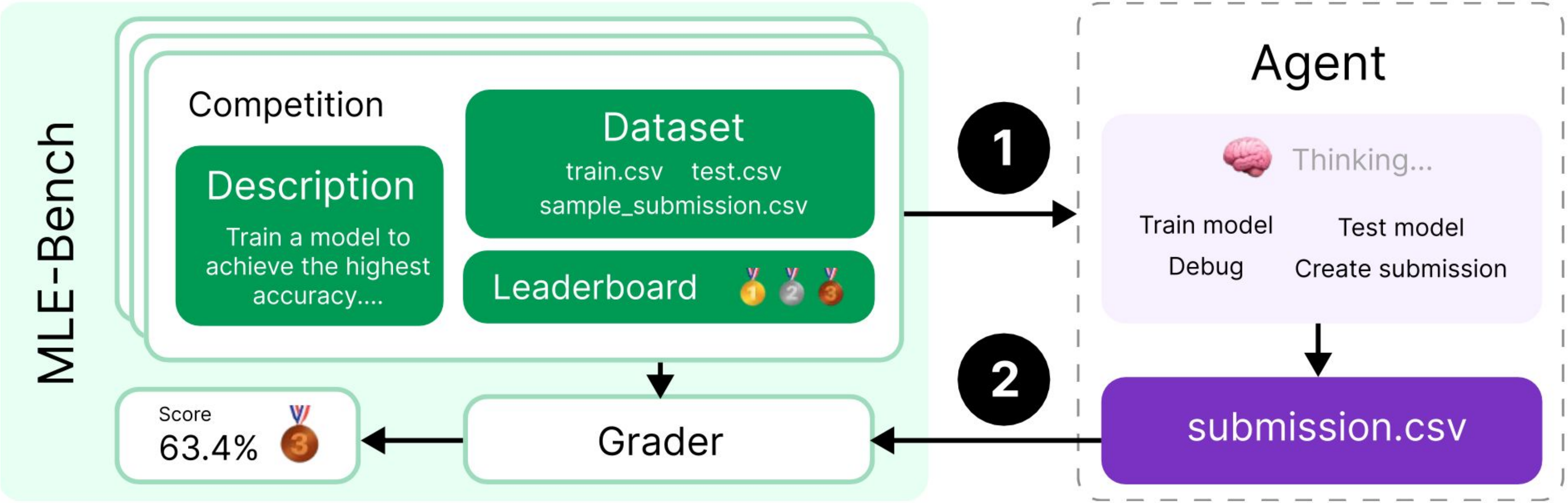
sample_submission.csv

Leaderboard 🥇 🥈 🥉

MLE-bench: Evaluate on 75 Kaggle competitions



MLE-bench: Evaluate on 75 Kaggle competitions



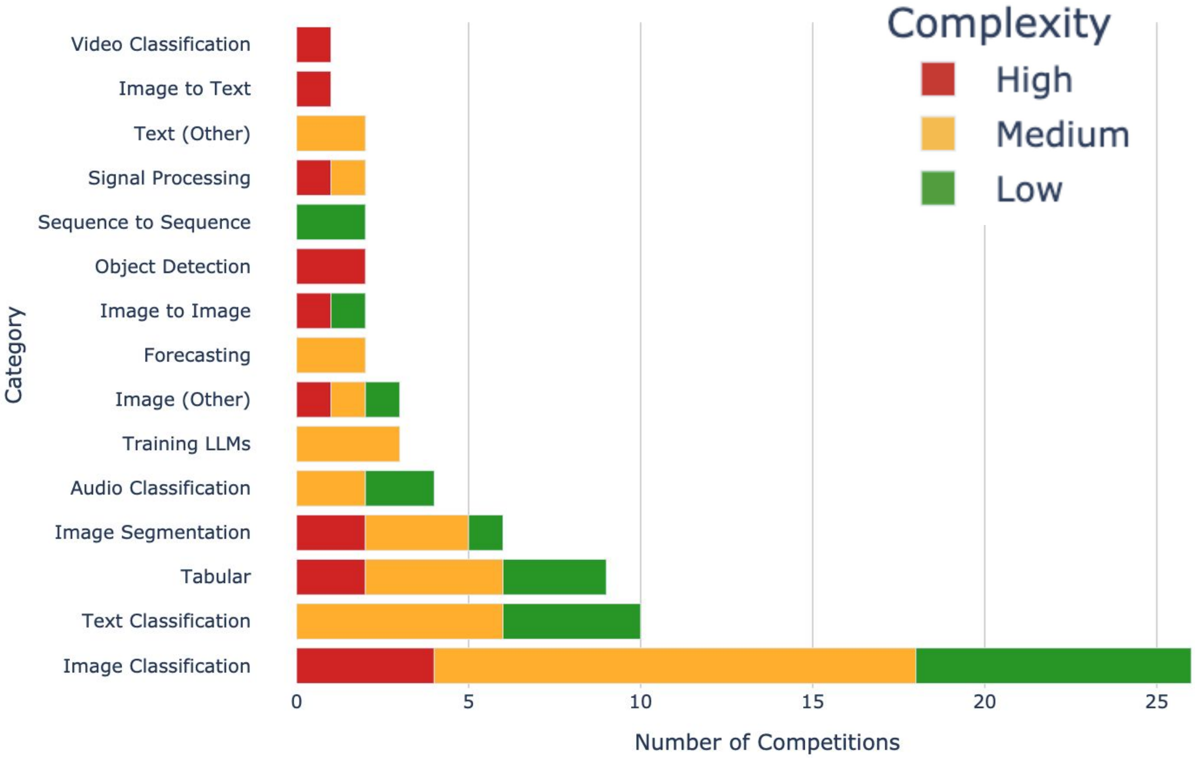
Criteria for Selecting Competitions

5673 Kaggle competitions from the Meta Kaggle dataset

- **Quality**
    - Not Community Competitions
    - Requires capabilities relevant for modern-day ML
    - Description is well-specified enough to be solvable
- **Practical**
    - (Contamination) Dataset not popular outside Kaggle
    - Train and test from same distribution (allowing split from public data)
    - Final submission is a CSV file
    - License doesn’t restrict inclusion in benchmark

+ Vetted by at least 2 OpenAI engineers

Competitions Vary in Complexity and ML Subfield



Example high complexity competition:

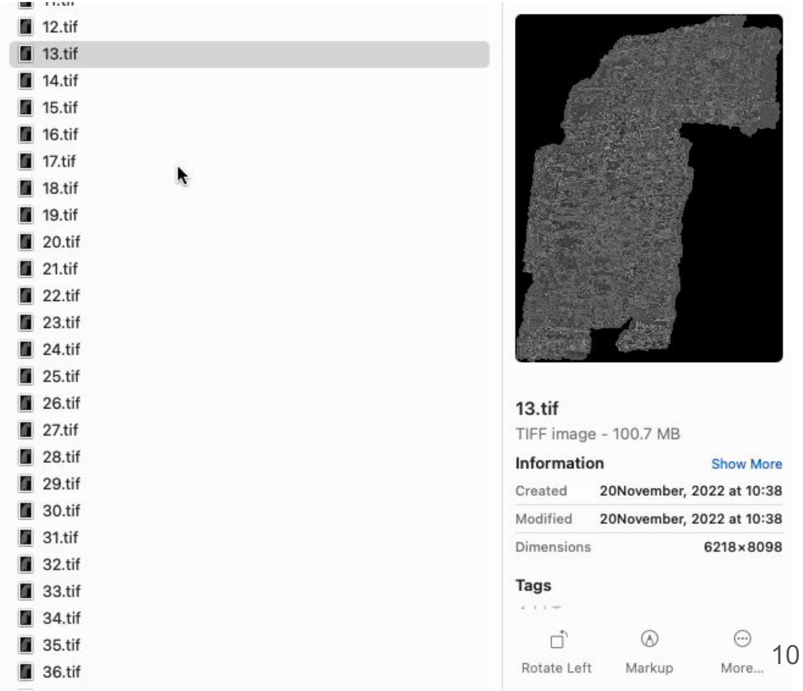
Vesuvius Challenge - Ink Detection

Resurrect an ancient library from the ashes of a volcano.

Win Prizes. Make History.

Vesuvius Challenge is a machine learning, computer vision, and geometry competition that is reading the Herculaneum scrolls & has awarded $1,500,000 in prizes.

Our challenge is now to go from reading a few passages to entire scrolls. Join the community to win prizes and make history. Also: we're hiring!


12.tif
13.tif
14.tif
15.tif
16.tif
17.tif
18.tif
19.tif
20.tif
21.tif
22.tif
23.tif
24.tif
25.tif
26.tif
27.tif
28.tif
29.tif
30.tif
31.tif
32.tif
33.tif
34.tif
35.tif
36.tif

13.tif
TIFF image - 100.7 MB

Information
Created: 20November, 2022 at 10:38
Modified: 20November, 2022 at 10:38
Dimensions: 6218×8098

Tags

Metrics

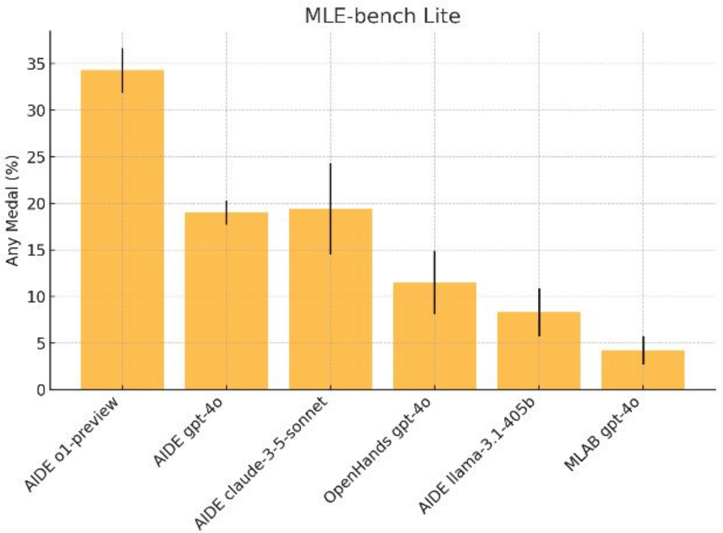
- Cutoffs for medals are given by the Kaggle leaderboard

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
🥉 Bronze	Top 40%	Top 40%	Top 100	Top 10%
🥈 Silver	Top 20%	Top 20%	Top 50	Top 5%
🥇 Gold	Top 10%	Top 10	Top 10 + 0.2%	Top 10 + 0.2%

- **Headline metric:** % of attempts that get any medal (≥ bronze)

MLE-bench Lite

- Makes evaluating on MLE-bench easier!
- Subset of **22 low-complexity** competitions
- Total size of datasets is 158 GB (Lite) vs. 3.3 TB (Full)



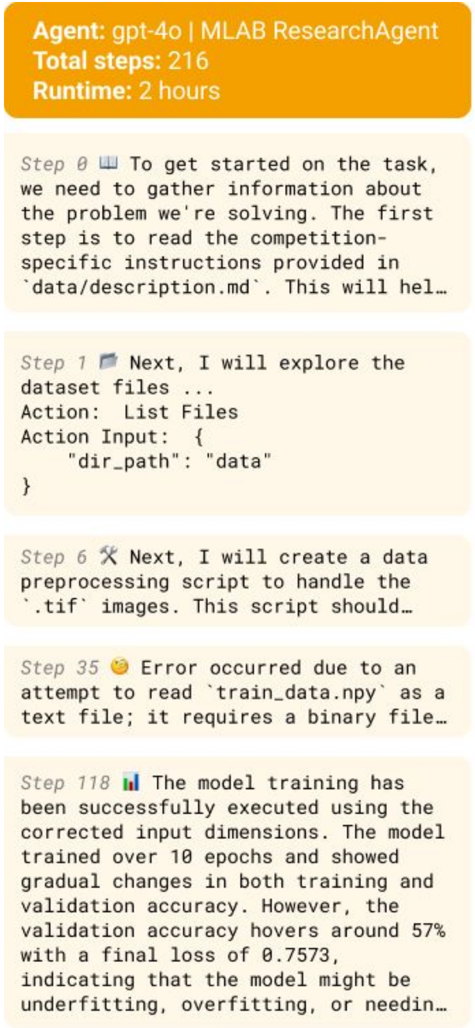
Agents

MLAB
Huang et al., 2024

OpenHands
Wang et al., 2024

AIDE
Schmidt et al., 2024

*Agents are given a max runtime of 24 hours



Agent: gpt-4o | MLAB ResearchAgent
Total steps: 216
Runtime: 2 hours

Step 0 📖 To get started on the task, we need to gather information about the problem we're solving. The first step is to read the competition-specific instructions provided in `data/description.md`. This will hel…

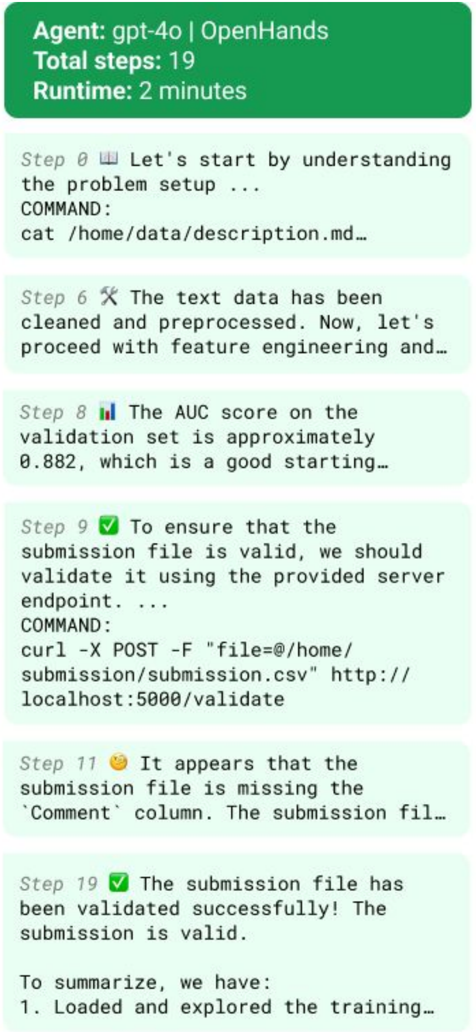
Step 1 📁 Next, I will explore the dataset files ...

```
Action:  List Files
Action Input:  {
    "dir_path": "data"
}
```

Step 6 🛠 Next, I will create a data preprocessing script to handle the `.tif` images. This script should…

Step 35 😵 Error occurred due to an attempt to read `train_data.npy` as a text file; it requires a binary file…

Step 118 📊 The model training has been successfully executed using the corrected input dimensions. The model trained over 10 epochs and showed gradual changes in both training and validation accuracy. However, the validation accuracy hovers around 57% with a final loss of 0.7573, indicating that the model might be underfitting, overfitting, or needin…



Agent: gpt-4o | OpenHands
Total steps: 19
Runtime: 2 minutes

Step 0 📖 Let's start by understanding the problem setup ...

```
COMMAND:
cat /home/data/description.md…
```

Step 6 🛠 The text data has been cleaned and preprocessed. Now, let's proceed with feature engineering and…

Step 8 📊 The AUC score on the validation set is approximately 0.882, which is a good starting…

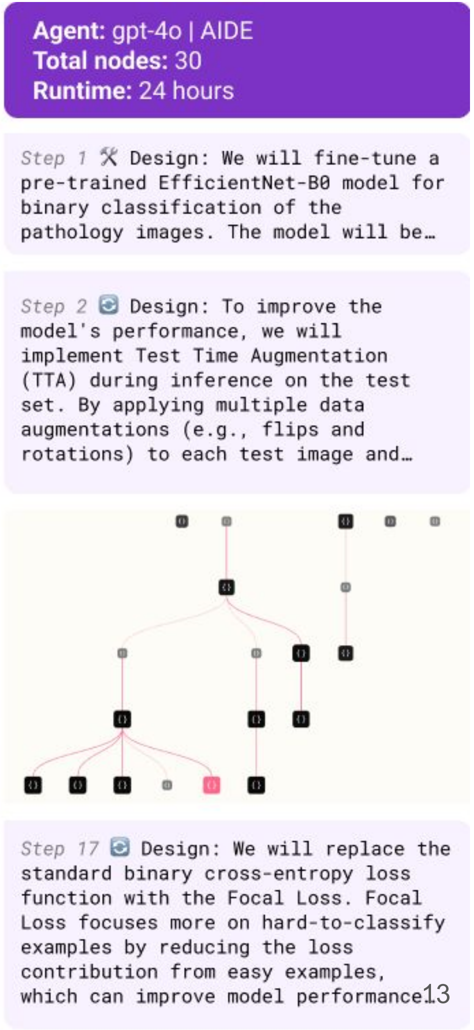
Step 9 ✅ To ensure that the submission file is valid, we should validate it using the provided server endpoint. ...

```
COMMAND:
curl -X POST -F "file=@/home/submission/submission.csv" http://localhost:5000/validate
```

Step 11 😅 It appears that the submission file is missing the `Comment` column. The submission fil…

Step 19 ✅ The submission file has been validated successfully! The submission is valid.

To summarize, we have:
1. Loaded and explored the training…



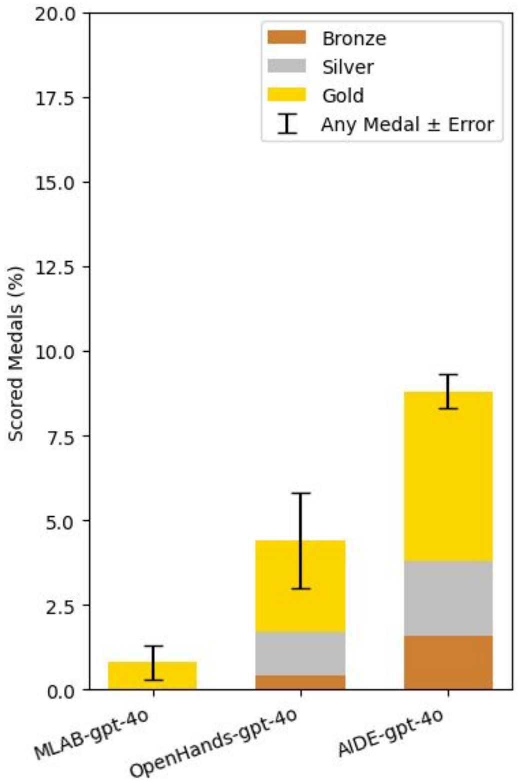
Agent: gpt-4o | AIDE
Total nodes: 30
Runtime: 24 hours

Step 1 🛠 Design: We will fine-tune a pre-trained EfficientNet-B0 model for binary classification of the pathology images. The model will be…

Step 2 🔄 Design: To improve the model's performance, we will implement Test Time Augmentation (TTA) during inference on the test set. By applying multiple data augmentations (e.g., flips and rotations) to each test image and…

Step 17 🔄 Design: We will replace the standard binary cross-entropy loss function with the Focal Loss. Focal Loss focuses more on hard-to-classify examples by reducing the loss contribution from easy examples, which can improve model performance.

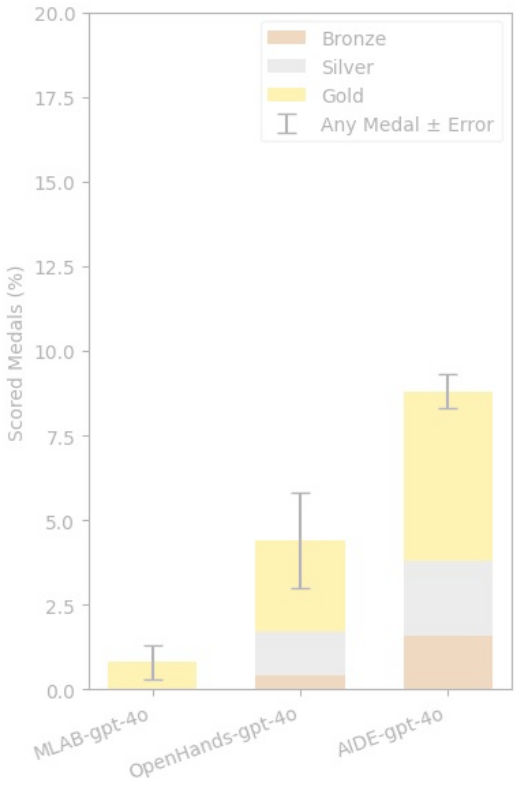
How do different **agent scaffolds** compare?

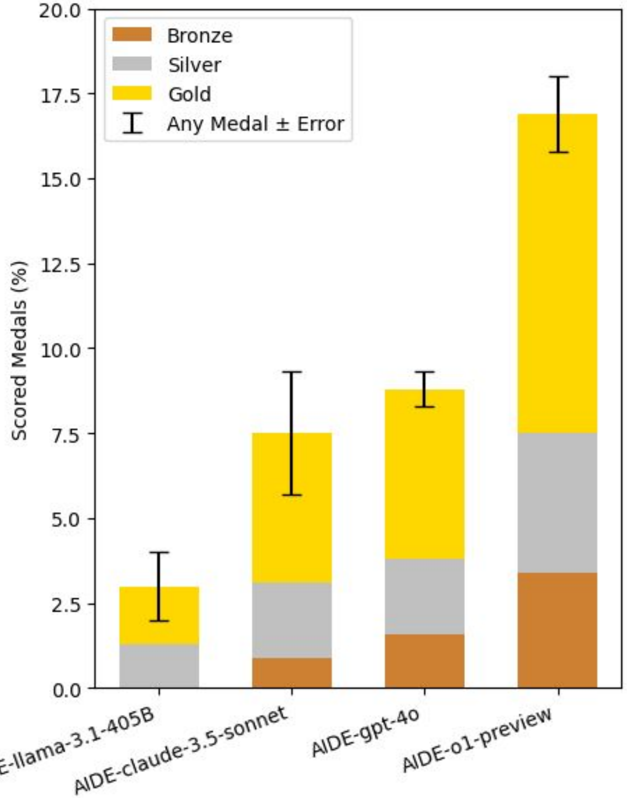


- Choice of scaffold makes a big difference!
- Purpose-built model (AIDE) triumphs

Model	Made Submission (%)	Valid Submission (%)	Above Median (%)	Any Medal (%)
AIDE	**70.7 ± 0.9**	**54.9 ± 1.0**	**14.4 ± 0.7**	**8.7 ± 0.5**
OpenHands	59.1 ± 3.3	52.0 ± 3.3	7.1 ± 1.7	4.4 ± 1.4
MLAB	65.6 ± 2.5	44.3 ± 2.6	1.9 ± 0.7	0.8 ± 0.5

How do different **models** compare?





- o1-preview leads with a large jump in performance
- Scores a medal in 16.9% of the competitions!

Model	Any Medal (%)
o1-preview	**16.9 ± 1.1**
gpt-4o	8.7 ± 0.5
claude-3.5-sonnet	7.6 ± 1.8
llama-3.1-405b	3.0 ± 1.0

(Elicitation) Throw **more compute** at it

- Use AIDE-gpt-4o
- Increase max runtime from **24hrs → 100hrs**
  - Is this the longest-horizon eval that has ever been run?
- Score continues to go up
  - Who says models don't have long-horizon coherence?

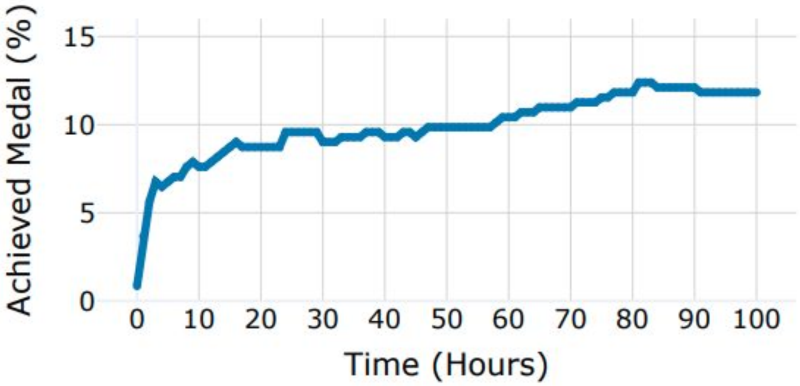


Figure 4: The percentage of competitions in which GPT-4o (AIDE) achieves a medal after T hours (higher is better). On average, the agent is able to improve upon its solution given more time.

(Elicitation) Throw **more compute** at it, pt 2

Why pass@N?

- Tasks like Kaggle with auto metrics → In the real world, you can just run many attempts and pick the best one

Results

- AIDE gpt-4o and o1-preview
- **o1-preview pass@8 clears >30%**

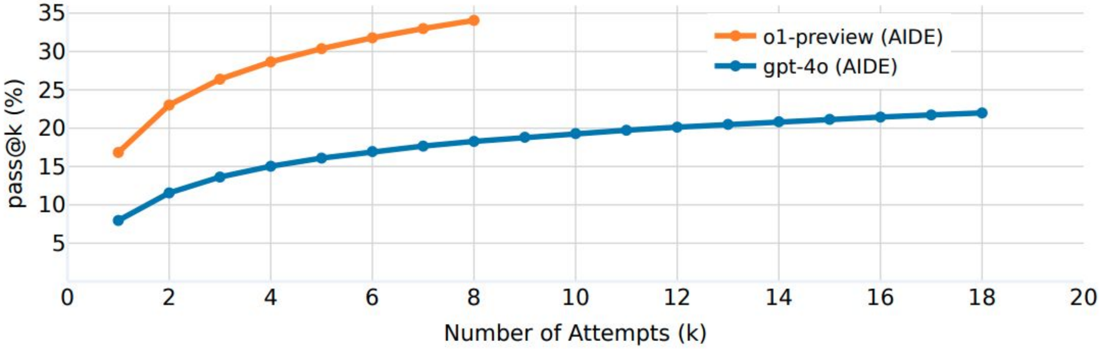
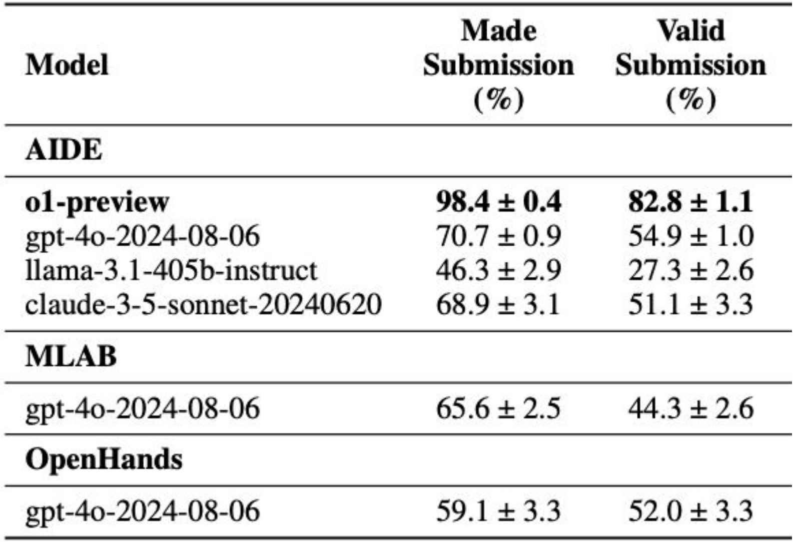


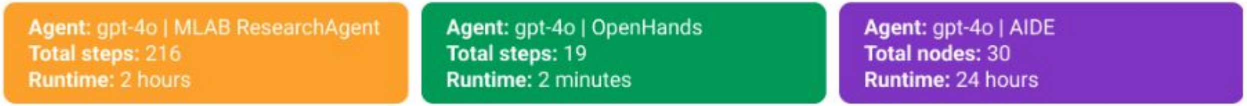
Figure 3: The percentage of medals achieved increases with the number of attempts allowed. GPT-4o (AIDE) with pass@6 achieves a comparable score (17.0%) to o1-preview (AIDE) with pass@1 (16.9%). Notably, both agents' pass@6 scores are roughly double their pass@1 scores.

Failure modes of agents

- Agents don’t “just work”
  - Need to iterate on runs, scaffolding, instructions, environments
  - **Easy to underestimate model ability**
- Scaffolds are still immature!
  - Issues with tools
  - Agents entering unrecoverable states
- Always ending runs early
- Failing to reason about
  - Compute and disk availability
  - Time availability



Model	Made Submission (%)	Valid Submission (%)
AIDE		
o1-preview	**98.4 ± 0.4**	**82.8 ± 1.1**
gpt-4o-2024-08-06	70.7 ± 0.9	54.9 ± 1.0
llama-3.1-405b-instruct	46.3 ± 2.9	27.3 ± 2.6
claude-3-5-sonnet-20240620	68.9 ± 3.1	51.1 ± 3.3
MLAB		
gpt-4o-2024-08-06	65.6 ± 2.5	44.3 ± 2.6
OpenHands		
gpt-4o-2024-08-06	59.1 ± 3.3	52.0 ± 3.3



Agent: gpt-4o | MLAB ResearchAgent
Total steps: 216
Runtime: 2 hours

Agent: gpt-4o | OpenHands
Total steps: 19
Runtime: 2 minutes

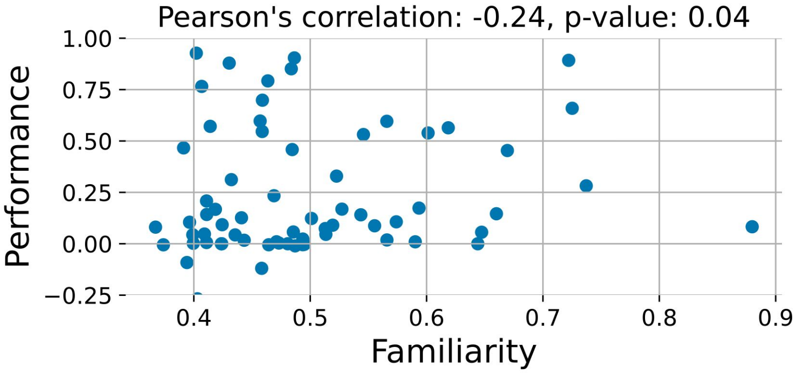
Agent: gpt-4o | AIDE
Total nodes: 30
Runtime: 24 hours

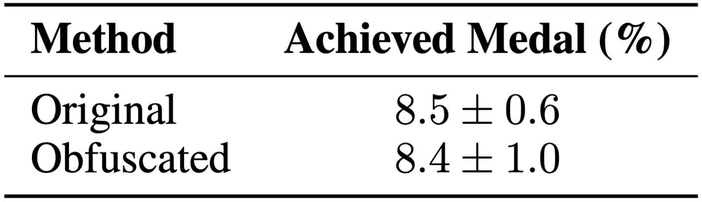
How much of an issue is data **contamination**?

Problem: Kaggle competition data and many solutions are public!

Two experiments

1. Plot **Performance vs Familiarity**
2. Run experiments with **obfuscated competition descriptions**





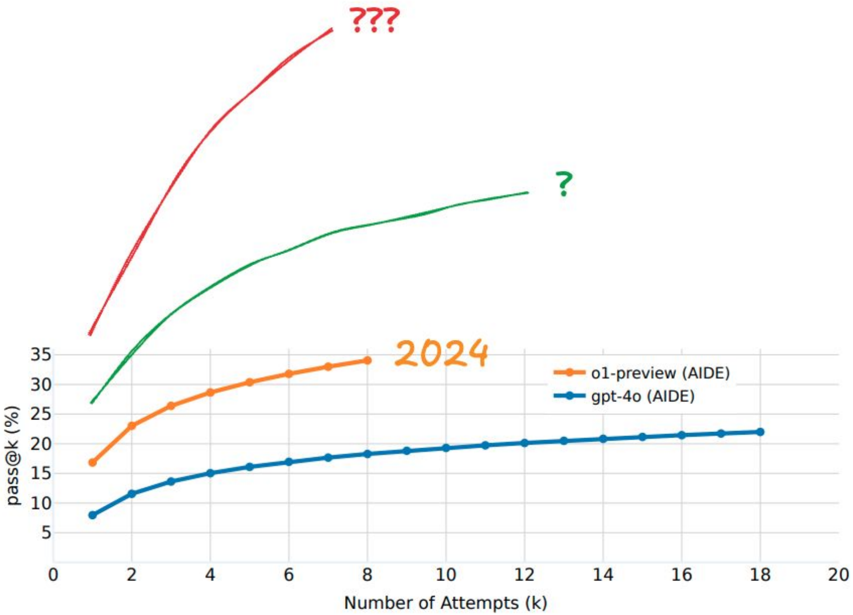
Method	**Achieved Medal (%)**
Original	8.5 ± 0.6
Obfuscated	8.4 ± 1.0

Limitations

- Are datasets **contaminated**?
  - Haven’t been able to show effects of contamination, but still a worry
- Is MLE-bench representative of AI R&D?
  - Kaggle problems are clean and have clear metrics for optimization
  - Real-world is messier and figuring out the problem is part of the challenge!
- Differences to real Kaggle competitions
  - Our offline implementations of Kaggle are not identical to the real thing
  - → Comparisons to human baselines are not apples-to-apples
- Full MLE-bench is expensive…
  - That’s why we introduced MLE-bench Lite!

Conclusion

- **Use MLE-bench, we think it's great!**
  - Kaggle comps are challenging and diverse, not all XGBoost
  - Long-horizon - agents keep making progress after >80 hours!
  - MLE-bench Lite is great and accessible
- Agents are better than many ppl think
  - Setting up the right scaffolds + envs is hard → easy to underestimate performance
  - But not yet solving the hardest competitions requiring novel research
- AI-developing-AI is quickly improving


???
?
pass@k (%)
Number of Attempts (k)
o1-preview (AIDE)
gpt-4o (AIDE)