AGI-Elo: How far are we from mastering a task?

Shuo Sun[1,3], Yimin Zhao[1], Christina Dao Wen Lee[1], Jiawei Sun[1], Chengran Yuan[1],
Zefan Huang[1,3], Dongen Li[1,3], Justin KW Yeoh[1], Alok Prakash[3],
Thomas W. Malone[2,3], Marcelo H. Ang Jr.[1,3]

[1]National University of Singapore, [2]Massachusetts Institute of Technology,
[3]Singapore MIT Alliance for Research and Technology

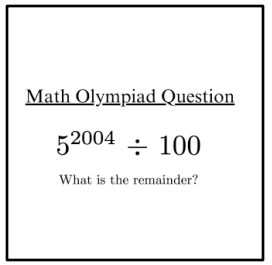
Motivation

Understanding Task Difficulty

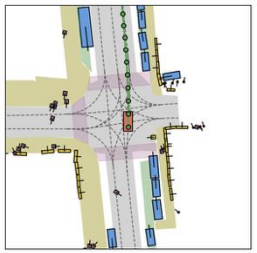
- How difficult is a task (to AI or Human)?

Playing Chess

Detecting Objects



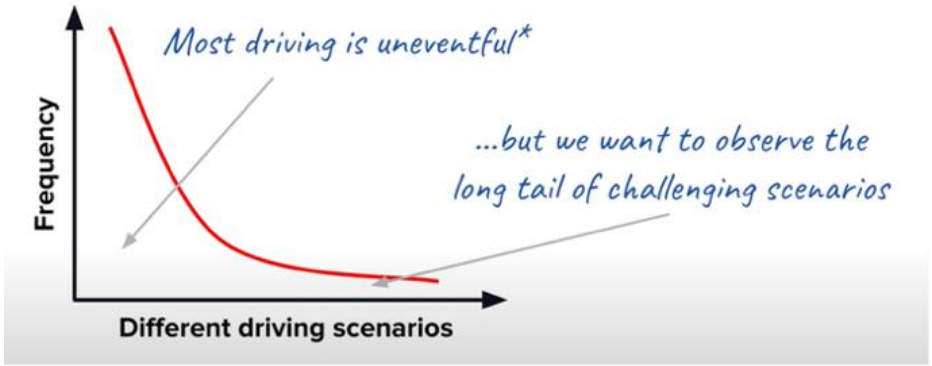
Answering Questions



Driving Vehicles

Tasks

- Have current SOTA models fully mastered a task?
  - Aggregate metrics ≠ Expected performance
  - Easy and difficult cases should not be treated equally
  - Data distribution is important
  - Quantifying the notion of difficulty is the key!

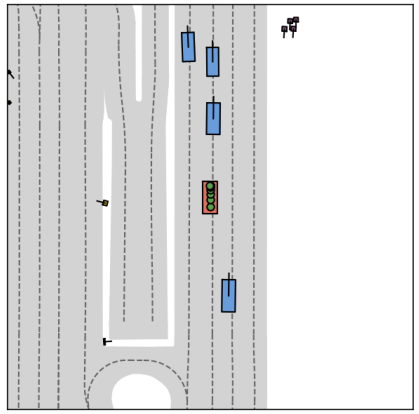
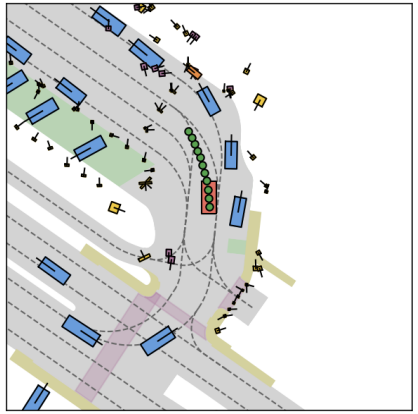
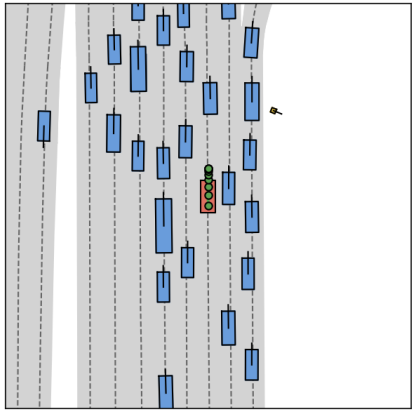
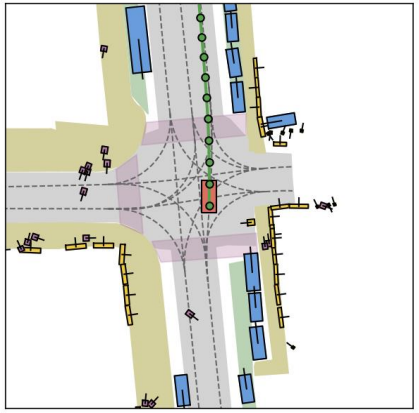


https://huggingface.co/datasets/ztony0712/motion_planning

Research Gap

Understanding Test Case Difficulty

- What is the difficulty of a test case within a task (or dataset)?
- What is the competency of an AI model (or a human) on a given task?
- How far are the current SOTA models from fully mastering a task?

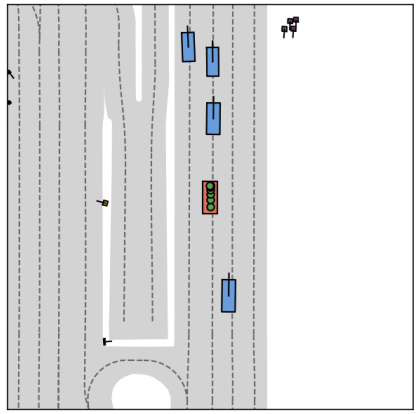
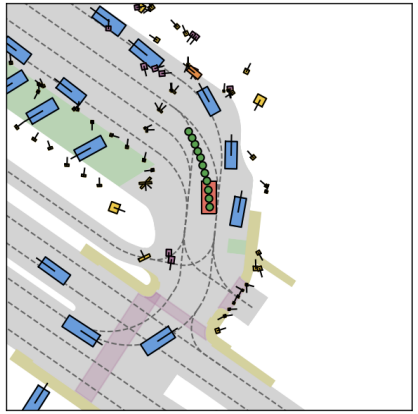
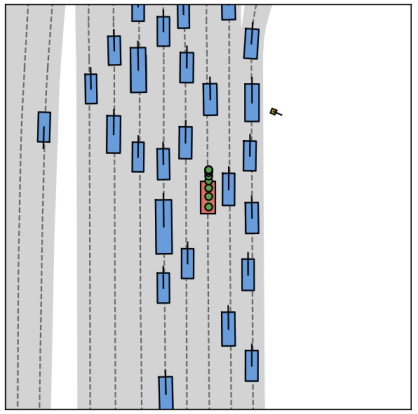
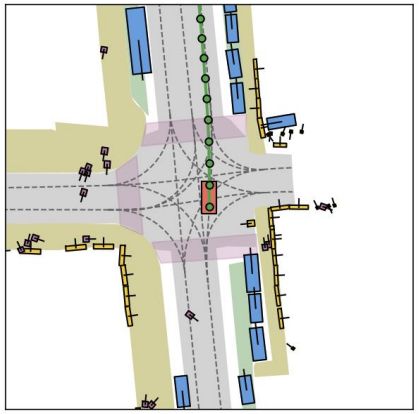
Difficulty ? ? ? ?

https://huggingface.co/datasets/ztony0712/motion_planning

Research Gap

Understanding Test Case Difficulty

- What is the difficulty of a test case within a task (or dataset)?
- What is the competency of an AI model (or a human) on a given task?
- How far are the current SOTA models from fully mastering a task?

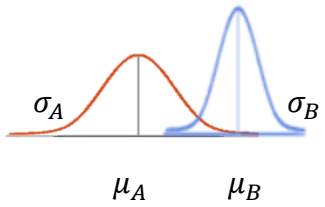
Difficulty	1099	1521	1875	2273
SOTA AI (2041)	99.44%	94.68%	75.97%	17.20%
Human Expert (2252)	99.87% ↑	98.68% ↑	93.02% ↑	46.69% ↑
Oracle@99% (3071)	100.00% ↑ ↑	99.99% ↑ ↑	99.90% ↑ ↑	99.00% ↑ ↑

https://huggingface.co/datasets/ztony0712/motion_planning

Rating System

AGI-Elo: Test Cases vs. Agents

- Jointly estimate the relative difficulty and competency:
  - Record model performance on each test case
  - Convert performance metric to match scores
  - Update ratings based on pairwise match outcomes
- Rating update (Glicko System):

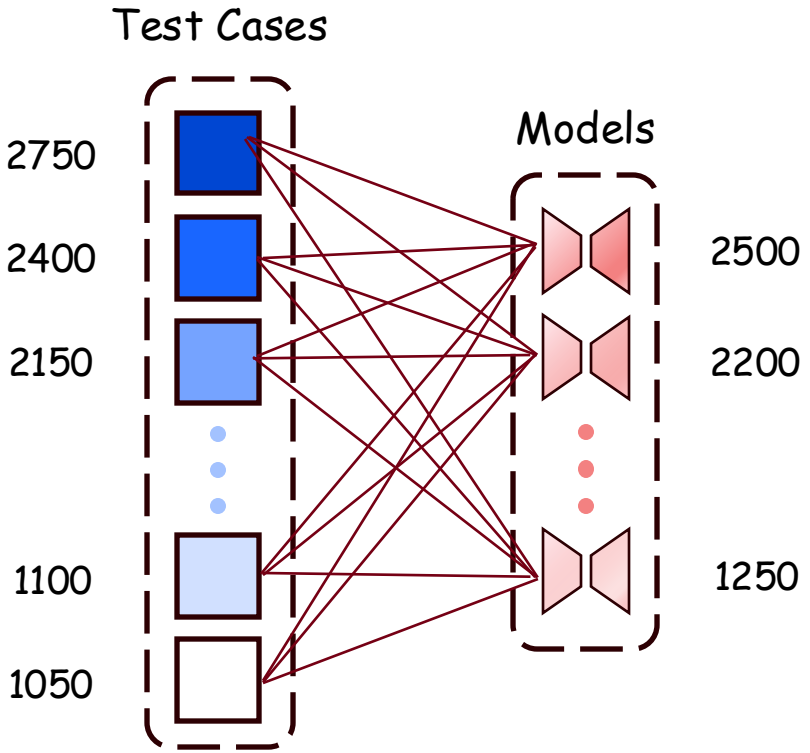


$$g(\sigma_j) = \frac{1}{\sqrt{1 + \frac{3q^2\sigma_j^2}{\pi^2}}}$$

$$E_{ij} = \frac{1}{1 + 10^{-g(\sigma_j)(\mu_i - \mu_j)/400}}$$

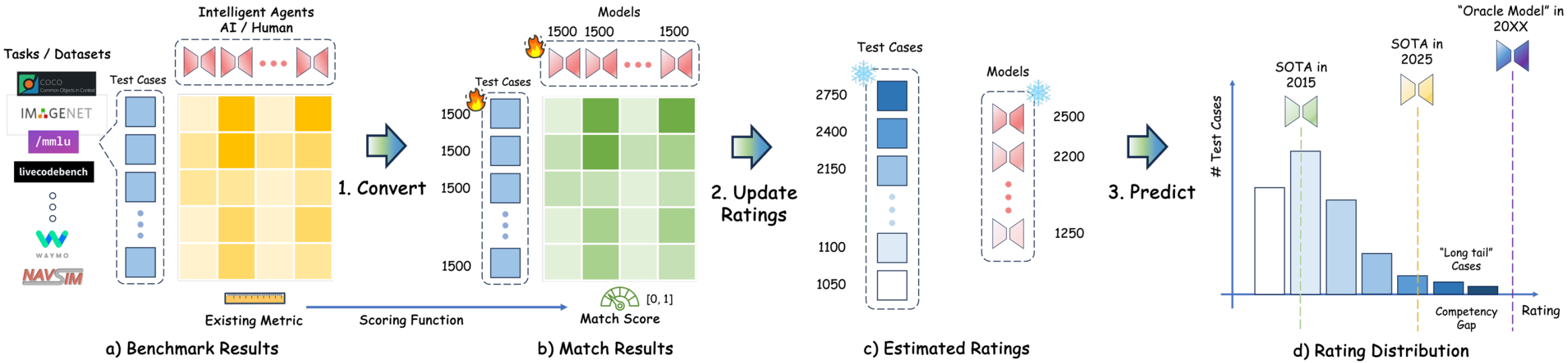
$$\mu_i \leftarrow \mu_i + \frac{q}{\frac{1}{\sigma_i^2} + \sum_j g(\sigma_j)^2 E_{ij}(1 - E_{ij})} \sum_j g(\sigma_j)(S_{ij} - E_{ij})$$

$$\sigma_i \leftarrow \left(\frac{1}{\sigma_i^2} + \sum_j g(\sigma_j)^2 E_{ij}(1 - E_{ij}) \right)^{-1/2}$$



Method

Test Case vs. Agents



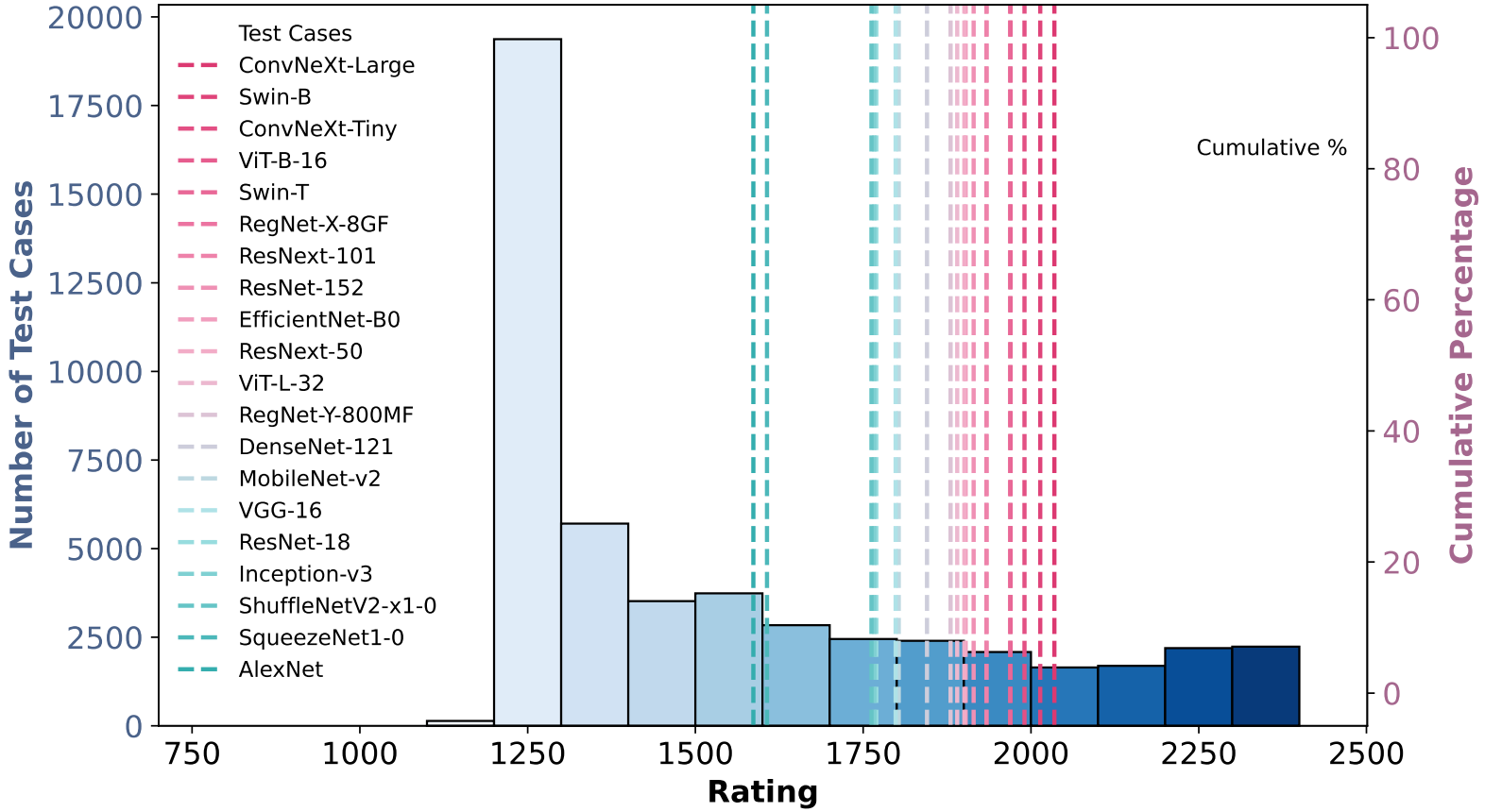
Findings

Rating Distribution

Vision Task:
- Image Classification

Dataset:
- ImageNet val set

Experiment:
- 5,000 test cases
- 20 models
- 1,000,000 matches



Qualitative Examples

Vision

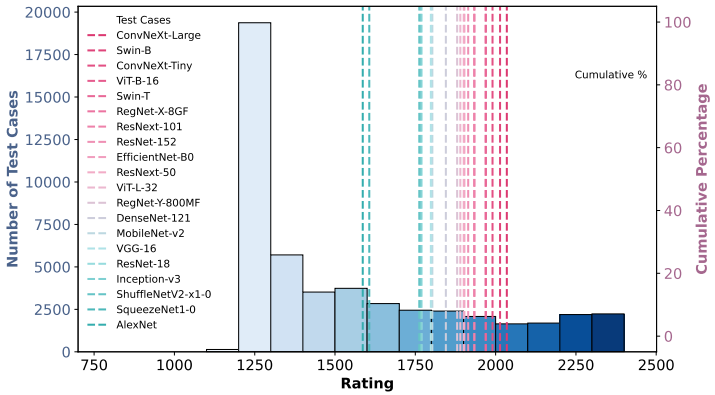
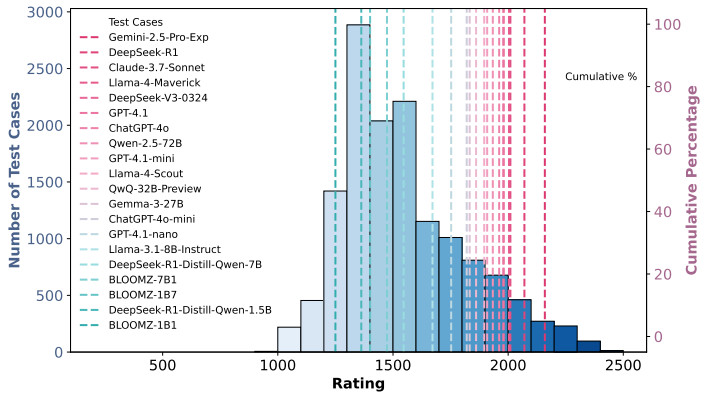


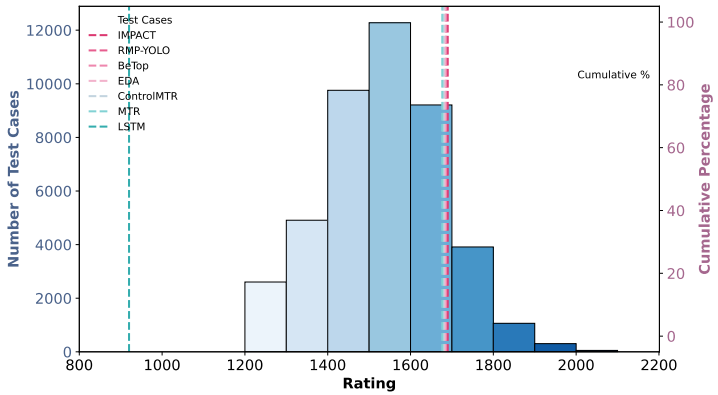
Image Classification: ImageNet

Language

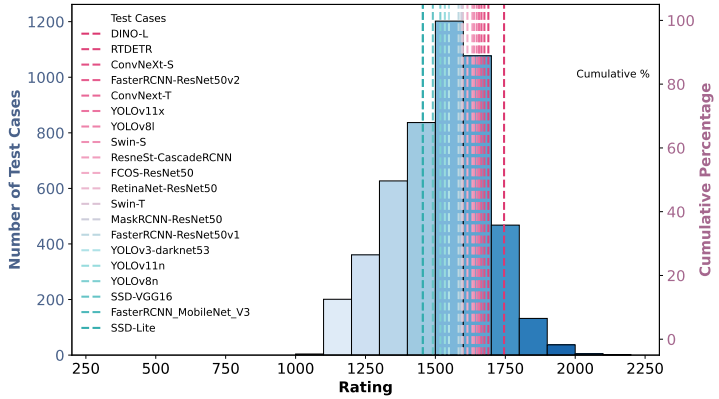


Question Answering: MMLU

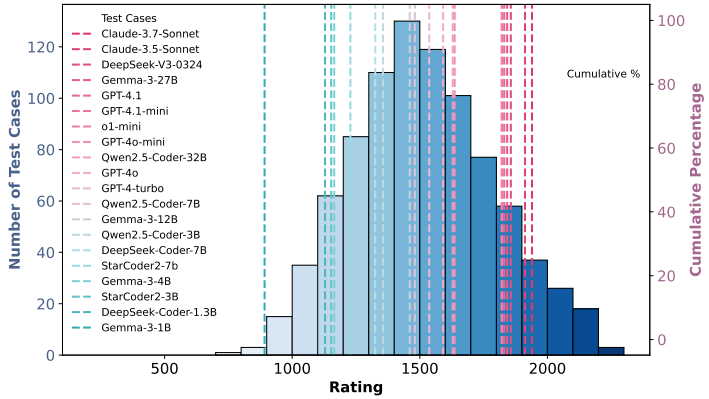
Action



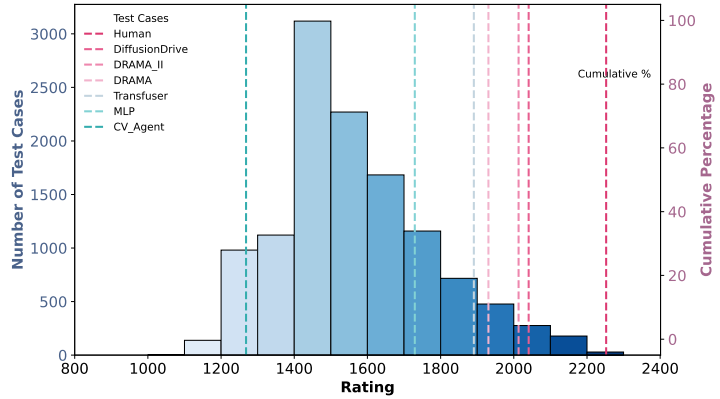
Motion Prediction: Waymo



Object Detection: COCO



Code Generation: LiveCodeBench



Motion Planning: NAVSIM

https://ss47816.github.io/AGI-Elo/

Conclusion

AGI-Elo provides a robust, predictive, and comprehensive framework for evaluating AGI capabilities and limitations.
By jointly modeling test case difficulty and model competency, we enable a deeper understanding of AI performance and trustworthiness across diverse domains.

Difficutly-aware Evaluation

Move beyond aggregate metrics to gain fine-grained insights into task difficulty distributions and model progression across vision, language, and action domains.

Predictive Insights

Quantitatively predict agent performance on individual test cases and identify long-tail challenges that require future progress for full task mastery.

Task-agnostic Benchmarking

Provide researchers and practitioners with a principled, task-agnostic framework for evaluating and comparing AI systems across diverse benchmarks.

Gap to Full Task Mastery

Identify outstanding challenges and competency gaps that remain on the path to achieving full task mastery at defined confidence levels.

Open Source & Community

Code and evaluated test case ratings on all six datasets are publicly available

https://ss47816.github.io/AGI-Elo/