September 12, 2024

Learning to reason with LLMs

We are introducing OpenAI o1, a new large language model trained with reinforcement learning to perform complex reasoning. o1 thinks before it answers —it can produce a long internal chain of thought before responding to the user.

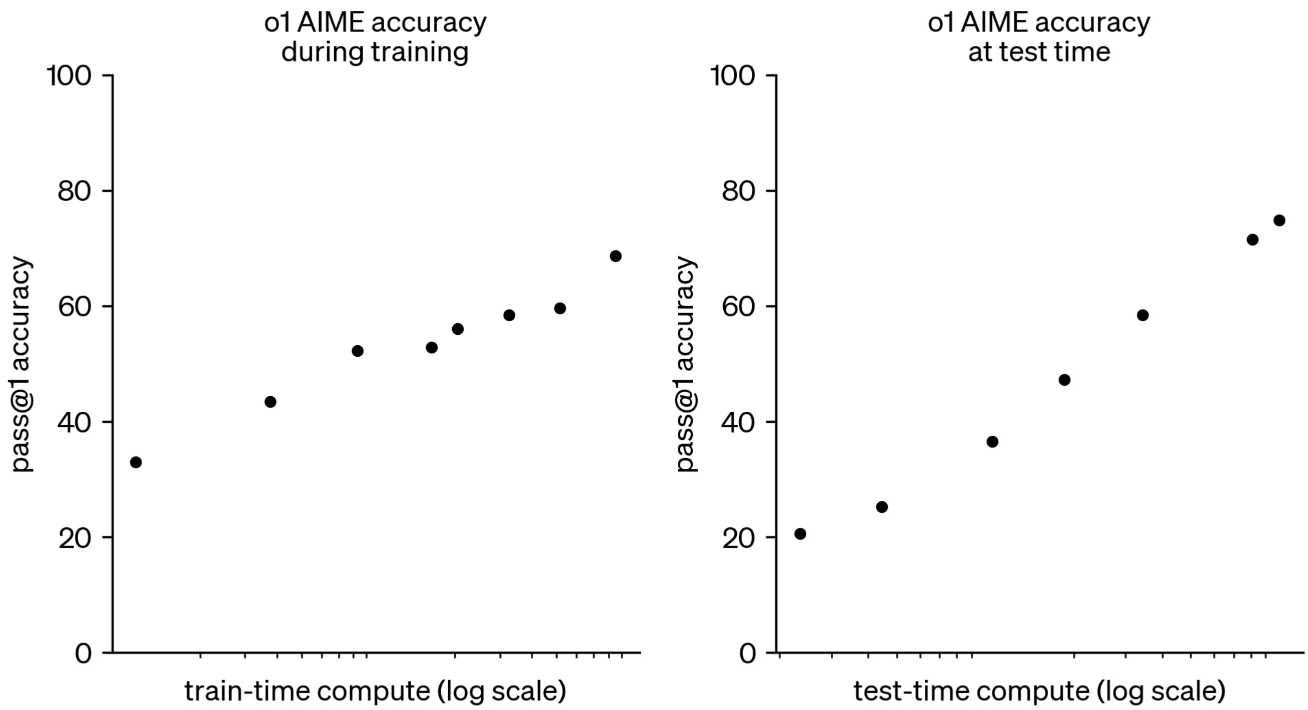
Contributions

Use o1 in ChatGPT Pro >

OpenAI o1 ranks in the 89th percentile on competitive programming questions (Codeforces), places among the top 500 students in the US in a qualifier for the USA Math Olympiad (AIME), and exceeds human PhD-level accuracy on a benchmark of physics, biology, and chemistry problems (GPQA). While the work needed to make this new model as easy to use as current models is still ongoing, we are releasing an early version of this model, OpenAI o1-preview, for immediate use in ChatGPT and to trusted API users.

Our large-scale reinforcement learning algorithm teaches the model how to think productively using its chain of thought in a highly data-efficient training process. We have found that the performance of o1 consistently improves with more reinforcement learning (train-time compute) and with more time spent thinking (test-time compute). The constraints on scaling this approach differ substantially from those of LLM pretraining, and we are continuing to investigate them.

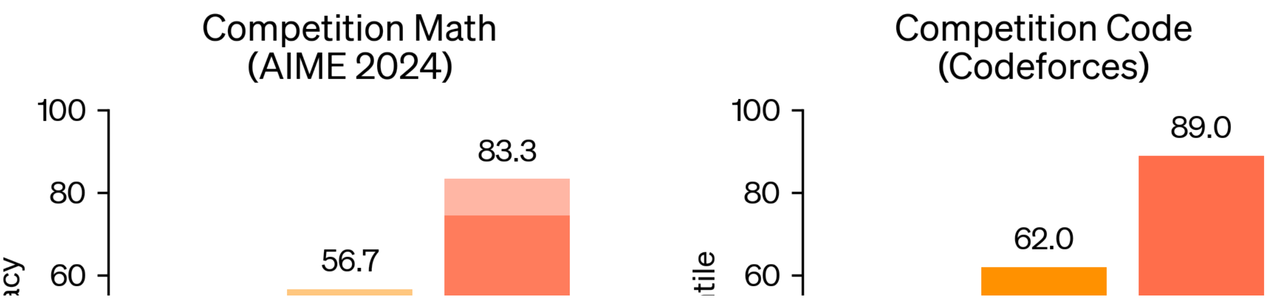
OpenAI

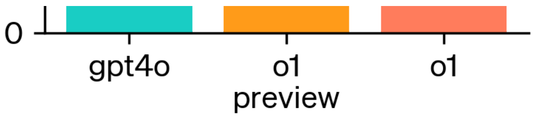


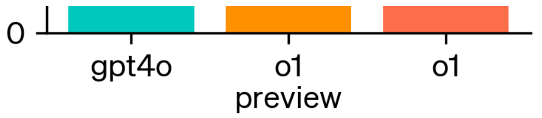
o1 performance smoothly improves with both train-time and test-time compute

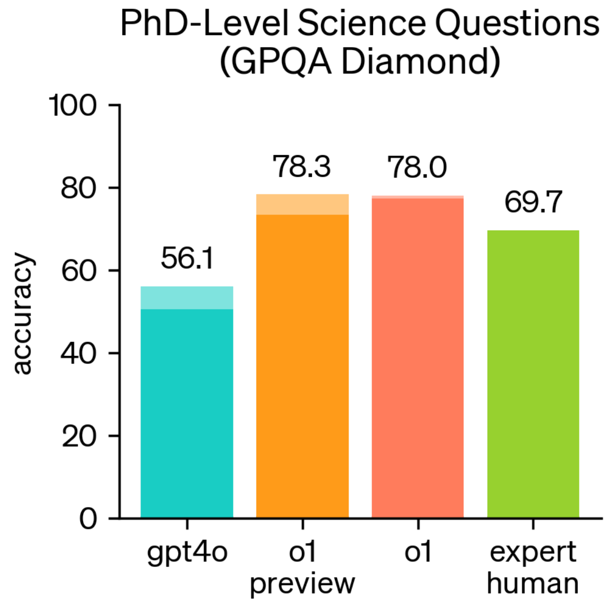
Evals

To highlight the reasoning improvement over GPT-4o, we tested our models on a diverse set of human exams and ML benchmarks. We show that o1 significantly outperforms GPT-4o on the vast majority of these reasoning-heavy tasks. Unless otherwise specified, we evaluated o1 on the maximal test-time compute setting.

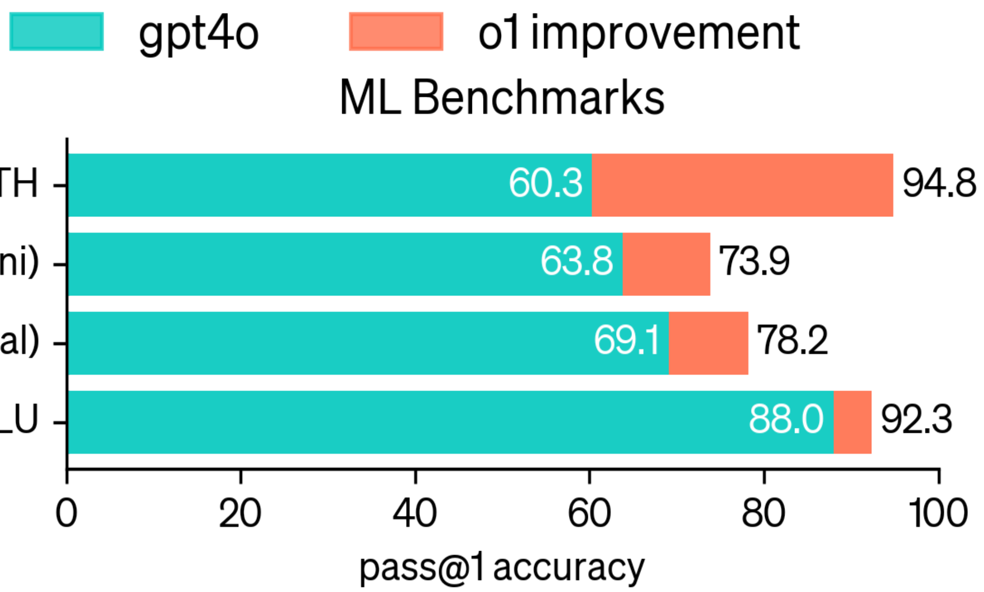


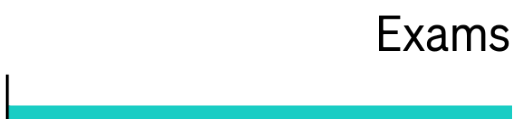


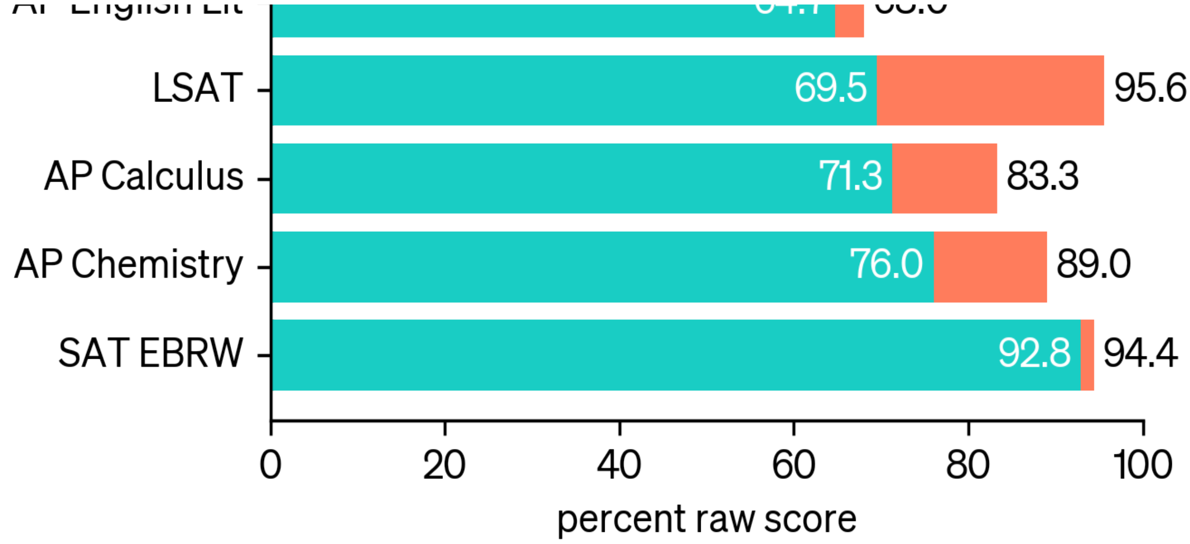




o1 greatly improves over GPT-4o on challenging reasoning benchmarks. Solid bars show pass@1 accuracy and the shaded region shows the performance of majority vote (consensus) with 64 samples.







o1 improves over GPT-4o on a wide range of benchmarks, including 54/57 MMLU subcategories. Seven are shown for illustration.

In many reasoning-heavy benchmarks, o1 rivals the performance of human experts. Recent frontier models[1] do so well on MATH[2] and GSM8K that these benchmarks are no longer effective at differentiating models. We evaluated math performance on AIME, an exam designed to challenge the brightest high school math students in America. On the 2024 AIME exams, GPT-4o only solved on average 12% (1.8/15) of problems. o1 averaged 74% (11.1/15) with a single sample per problem, 83% (12.5/15) with consensus among 64 samples, and 93% (13.9/15) when re-ranking 1000 samples with a learned scoring function. A score of 13.9 places it among the top 500 students nationally and above the cutoff for the USA Mathematical Olympiad.

We also evaluated o1 on GPQA diamond, a difficult intelligence benchmark which tests for expertise in chemistry, physics and biology. In order to compare models to humans, we recruited experts with PhDs to answer GPQA-diamond questions. We found that o1 surpassed the performance of those human experts, becoming the first model to do so on this benchmark. These results do not imply that o1 is more capable than a PhD in all respects — only that the model is more proficient in solving some problems that a PhD would be expected to solve. On several other ML benchmarks, o1 improved over the state-of-the-art. With its vision perception capabilities enabled, o1 scored 78.2% on MMMU, making it the first model to be competitive with human

Chain of Thought

Similar to how a human may think for a long time before responding to a difficult question, o1 uses a chain of thought when attempting to solve a problem. Through reinforcement learning, o1 learns to hone its chain of thought and refine the strategies it uses. It learns to recognize and correct its mistakes. It learns to break down tricky steps into simpler ones. It learns to try a different approach when the current one isn't working. This process dramatically improves the model's ability to reason. To illustrate this leap forward, we showcase the chain of thought from o1-preview on several difficult problems below.

r Coding Math Crossword English Science Safety Health Sc

GPT-4o

User

oyfjdnisdr rtqwainr acxz mynzbhhx -> Think step by step

Use the example above to decode:

oyekaijzdf aaptcg suaokybhai ouow aqht mynznvaatzacdfoulxxz

Let's break this down step by step based on the example:

1. Example given:

- Input: oyfjdnisdr rtqwainr acxz mynzbhhx
- Output: **Think step by step**

By examining the words:

- The pattern involves selecting specific letters or transforming them

OpenAI o1-preview

OpenAI

User

oyfjdnisdr rtqwainr acxz mynzbhhx -> Think step by step

Use the example above to decode:

oyekaijzdf aaptcg suaokybhai ouow aqht mynznvaatzacdfoulxxz

Thought for 5 seconds

To decode the given ciphertext using the example provided, we'll use the pattern identified in the example.

Ciphertext Example:

Coding

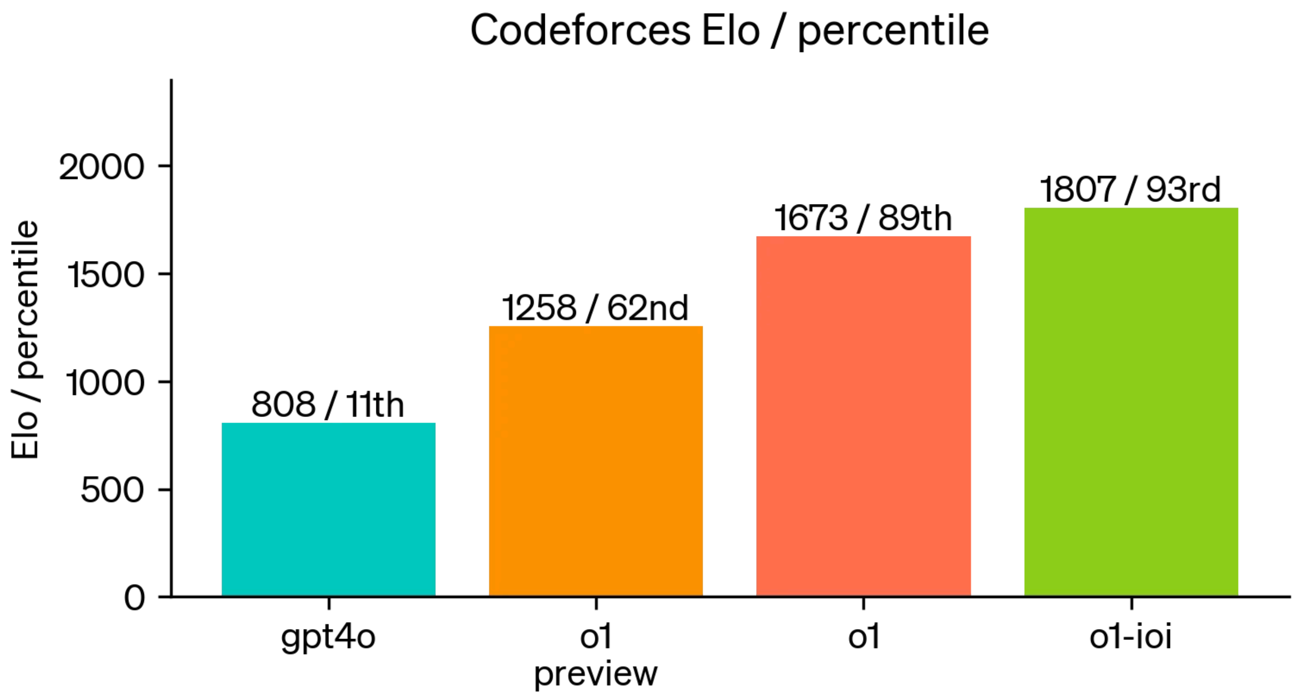
We trained a model that scored 213 points and ranked in the 49th percentile in the 2024 International Olympiad in Informatics (IOI), by initializing from o1 and training to further improve programming skills. This model competed in the 2024 IOI under the same conditions as the human contestants. It had ten hours to solve six challenging algorithmic problems and was allowed 50 submissions per problem.

For each problem, our system sampled many candidate submissions and submitted 50 of them based on a test-time selection strategy. Submissions were selected based on performance on the IOI public test cases, model-generated test cases, and a learned scoring function. If we had instead submitted at random, we would have only scored 156 points on average, suggesting that this strategy was worth nearly 60 points under competition constraints.

With a relaxed submission constraint, we found that model performance improved significantly. When allowed 10,000 submissions per problem, the model achieved a score of 362.14 – above the gold medal threshold – even without any test-time selection strategy.

OpenAI

demonstrate this model's coding skill. Our evaluations closely matched competition rules and allowed for 10 submissions. GPT-4o achieved an Elo rating[3] of 808, which is in the 11th percentile of human competitors. This model far exceeded both GPT-4o and o1—it achieved an Elo rating of 1807, performing better than 93% of competitors.



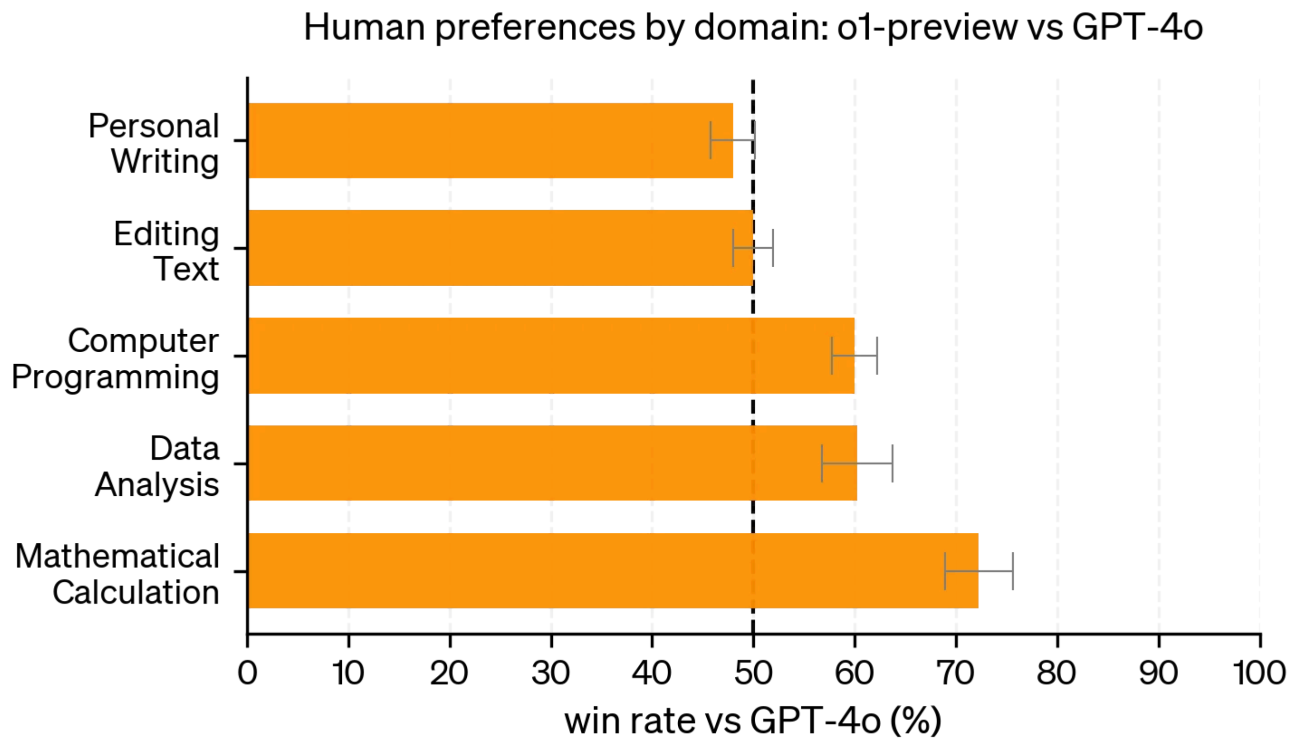
Further fine-tuning on programming competitions improves o1. The improved model ranked in the 49th percentile in the 2024 International Olympiad in Informatics under competition rules.

+

Human preference evaluation

In addition to exams and academic benchmarks, we also evaluated human preference of o1-preview vs GPT-4o on challenging, open-ended prompts in a broad spectrum of domains. In this evaluation, human trainers were shown anonymized responses to a prompt from o1-preview and GPT-4o, and voted for which response they preferred. o1-preview is preferred to gpt-4o by a large margin in reasoning-heavy categories like data analysis, coding, and math. However, o1-preview is not

OpenAI



Safety

Coding

Chain of thought reasoning provides new opportunities for alignment and safety. We found that integrating our policies for model behavior into the chain of thought of a reasoning model is an effective way to robustly teach human values and principles. By teaching the model our safety rules and how to reason about them in context, we found evidence of reasoning capability directly benefiting model robustness: o1-preview achieved substantially improved performance on key jailbreak evaluations and our hardest internal benchmarks for evaluating our model's safety refusal boundaries. We believe that using a chain of thought offers significant advances for safety and alignment because (1) it enables us to observe the model thinking in a legible way, and (2) the model reasoning about safety rules is more robust to out-of-distribution scenarios.

To stress-test our improvements, we conducted a suite of safety tests and red-teaming before deployment, in accordance with our Preparedness Framework. We

OpenAI

our evaluations. Of particular note, we observed interesting instances of reward hacking. Detailed results from these evaluations can be found in the accompanying System Card.

Metric	`GPT-4o`	`o1-preview`
% Safe completions on harmful prompts Standard	0.990	0.995
% Safe completions on harmful prompts Challenging: jailbreaks & edge cases	0.714	0.934
↳ Harassment (severe)	0.845	0.900
↳ Exploitative sexual content	0.483	0.949
↳ Sexual content involving minors	0.707	0.931
↳ Advice about non-violent wrongdoing	0.688	0.961
↳ Advice about violent wrongdoing	0.778	0.963
% Safe completions for top 200 with highest Moderation API scores per category in WildChat Zhao, et al. 2024	0.945	0.971
Goodness@0.1 StrongREJECT jailbreak eval Souly et al. 2024	0.220	0.840
Human sourced jailbreak eval	0.770	0.960
% Compliance on internal benign edge cases “not over-refusal”	0.910	0.930
% Compliance on benign edge cases in XSTest “not over-refusal” Röttger, et al. 2023	0.924	0.976

Hiding the Chains of Thought

OpenAI

We believe that a hidden chain of thought presents a unique opportunity for monitoring models. Assuming it is faithful and legible, the hidden chain of thought allows us to "read the mind" of the model and understand its thought process. For example, in the future we may wish to monitor the chain of thought for signs of manipulating the user. However, for this to work the model must have freedom to express its thoughts in unaltered form, so we cannot train any policy compliance or user preferences onto the chain of thought. We also do not want to make an unaligned chain of thought directly visible to users.

Therefore, after weighing multiple factors including user experience, competitive advantage, and the option to pursue the chain of thought monitoring, we have decided not to show the raw chains of thought to users. We acknowledge this decision has disadvantages. We strive to partially make up for it by teaching the model to reproduce any useful ideas from the chain of thought in the answer. For the o1 model series we show a model-generated summary of the chain of thought.

Conclusion

o1 significantly advances the state-of-the-art in AI reasoning. We plan to release improved versions of this model as we continue iterating. We expect these new reasoning capabilities will improve our ability to align models to human values and principles. We believe o1 – and its successors – will unlock many new use cases for AI in science, coding, math, and related fields. We are excited for users and API developers to discover how it can improve their daily work.

Appendix A

Dataset	Metric	`gpt-4o`	`o1-preview`	`o1`
Competition Math AIME (2024)	cons@64	13.4	56.7	83.3
	pass@1	9.3	44.6	74.4
Competition Code CodeForces	Elo	808	1,258	1,673

Dataset	Metric	`gpt-4o`	`o1-preview`	`o1`
GPQA Diamond	cons@64	56.1	78.3	78.0

Dataset	Metric	`gpt-4o`	`o1-preview`	`o1`
	pass@1	50.6	73.3	77.3
Biology	cons@64	63.2	73.7	68.4
	pass@1	61.6	65.9	69.2
Chemistry	cons@64	43.0	60.2	65.6
	pass@1	40.2	59.9	64.7
Physics	cons@64	68.6	89.5	94.2
	pass@1	59.5	89.4	92.8
MATH	pass@1	60.3	85.5	94.8
MMLU	pass@1	88.0	90.8	92.3
MMMU (val)	pass@1	69.1	n/a	78.2
MathVista (testmini)	pass@1	63.8	n/a	73.9

Authors

OpenAI

View contributors >

Citations

OpenAI

2 Our evaluations used the same 500 problem test split found in https://arxiv.org/abs/2305.20050 ↩

3 https://codeforces.com/blog/entry/68288 ↩

API

Platform overview

Pricing

Documentation ↗

API login ↗

Explore more

OpenAI for business

Stories

Safety overview

Safety overview

Company

About us

News

Our Charter

Security

Residency

Careers

Terms & policies

Terms of use

Privacy policy

Brand guidelines

Other policies

OpenAI

English (US)