April 14, 2025 Product

Introducing GPT-4.1 in the API

A new series of GPT models featuring major improvements on coding, instruction following, and long context—plus our first-ever nano model.

Try in Playground

Listen to article 18:10 Share

Today, we're launching three new models in the API: GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano. These models outperform GPT-4o and GPT-4o mini across the board, with major gains in coding and instruction following. They also have larger context windows—supporting up to 1 million tokens of context—and are able to better use that context with improved long-context comprehension. They feature a refreshed knowledge cutoff of June 2024.

GPT-4.1 excels at the following industry standard measures:

- **Coding**: GPT-4.1 scores 54.6% on SWE-bench Verified, improving by *21.4%$_{abs}$* over GPT-4o *and 26.6%$_{abs}$* over GPT-4.5—making it a leading model for coding.
- **Instruction following:** On Scale's MultiChallenge benchmark, a measure of instruction following ability, GPT-4.1 scores 38.3%, a 10.5%$_{\text{abs}}$ increase over GPT-4o.
- **Long context:** On Video-MME, a benchmark for multimodal long context understanding, GPT-4.1 sets a new state-of-the-art result—scoring 72.0% on the long, no subtitles category, a 6.7%$_{\text{abs}}$ improvement over GPT-4o.

2025年4月14日 产品

在API中引入GPT-4.1

一系列新的GPT模型，在编码、指令遵循和长上下文方面实现了重大改进——以及我们首次推出的纳米模型。

在 Playground 中试用

听文章 18:10 Share

今天，我们在API中推出了三款新模型：GPT-4.1、GPT-4.1 mini 和 GPT-4.1 nano。这些模型在各方面都优于 GPT-4o 和 GPT-4o mini，在编码和指令执行方面取得了显著提升。它们还拥有更大的上下文窗口——支持高达 1 百万 tokens 的上下文——并且能够通过改进的长上下文理解更好地利用这些上下文。它们的知识截止日期也更新为 2024 年 6 月。

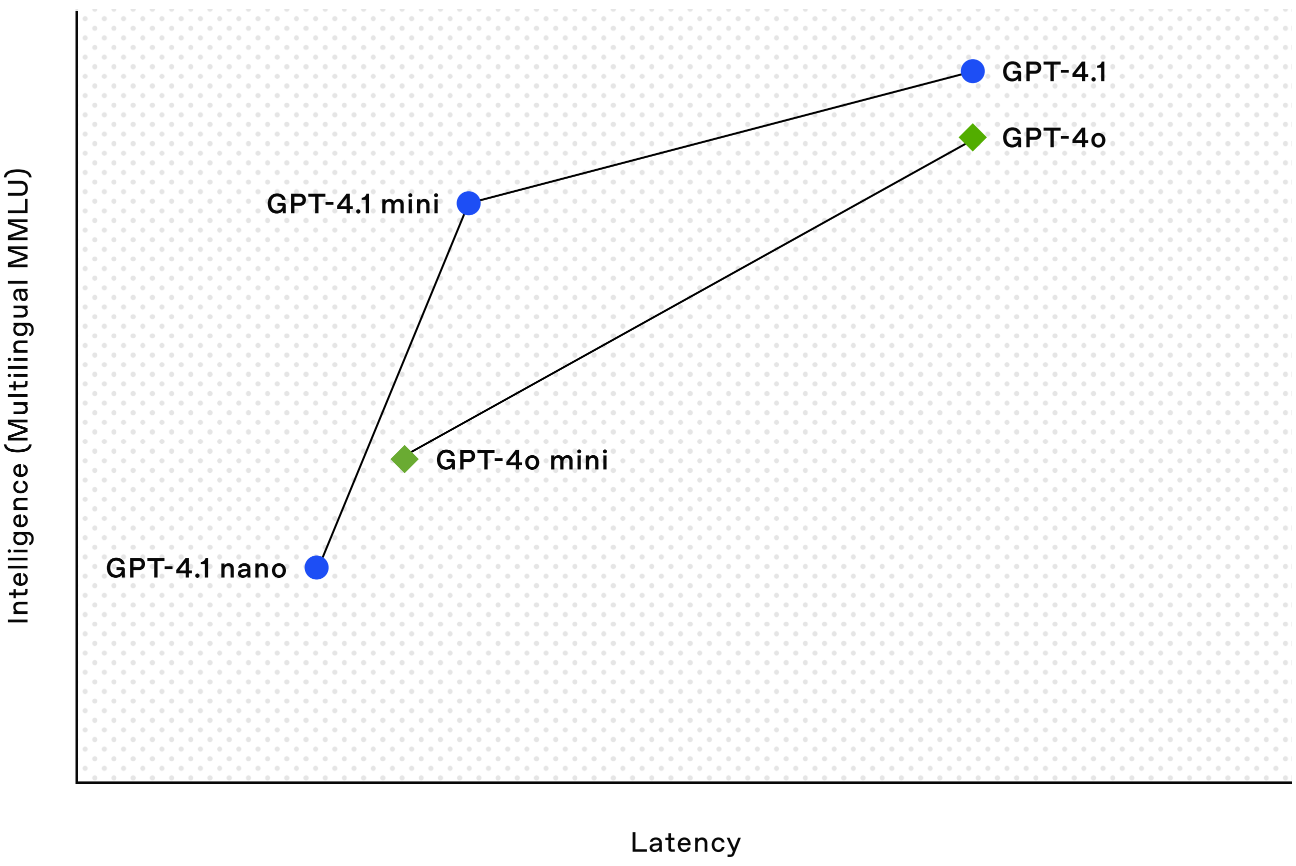
GPT-4.1 在以下行业标准指标方面表现出色：

- 编码：GPT-4.1 在 上得分为 54.6%，SWE-bench Verified 比 GPT-4o 提升了 abs abs *21.4%* ，比 GPT-4.5 提升了 *and 26.6%* —使其成为编码领域的领先模型。
- 在 Scale 的 MultiChallenge 基准测试中，衡量指令执行能力，GPT-4.1 的得分为 38.3%，比 GPT-4o 提升了 10.5%。 绝对值
- 长上下文：在 Video-MME 这一多模态长上下文理解的基准测试中，GPT-4.1 取得了新的最先进的成果——在长时间无字幕类别中得分为 72.0%，比 GPT-4o 提升了 6.7%。 绝对值

community enabled us to optimize these models for the tasks that matter most to their applications.

To this end, the GPT-4.1 model family offers exceptional performance at a lower cost. These models push performance forward at every point on the latency curve.

GPT-4.1 family intelligence by latency



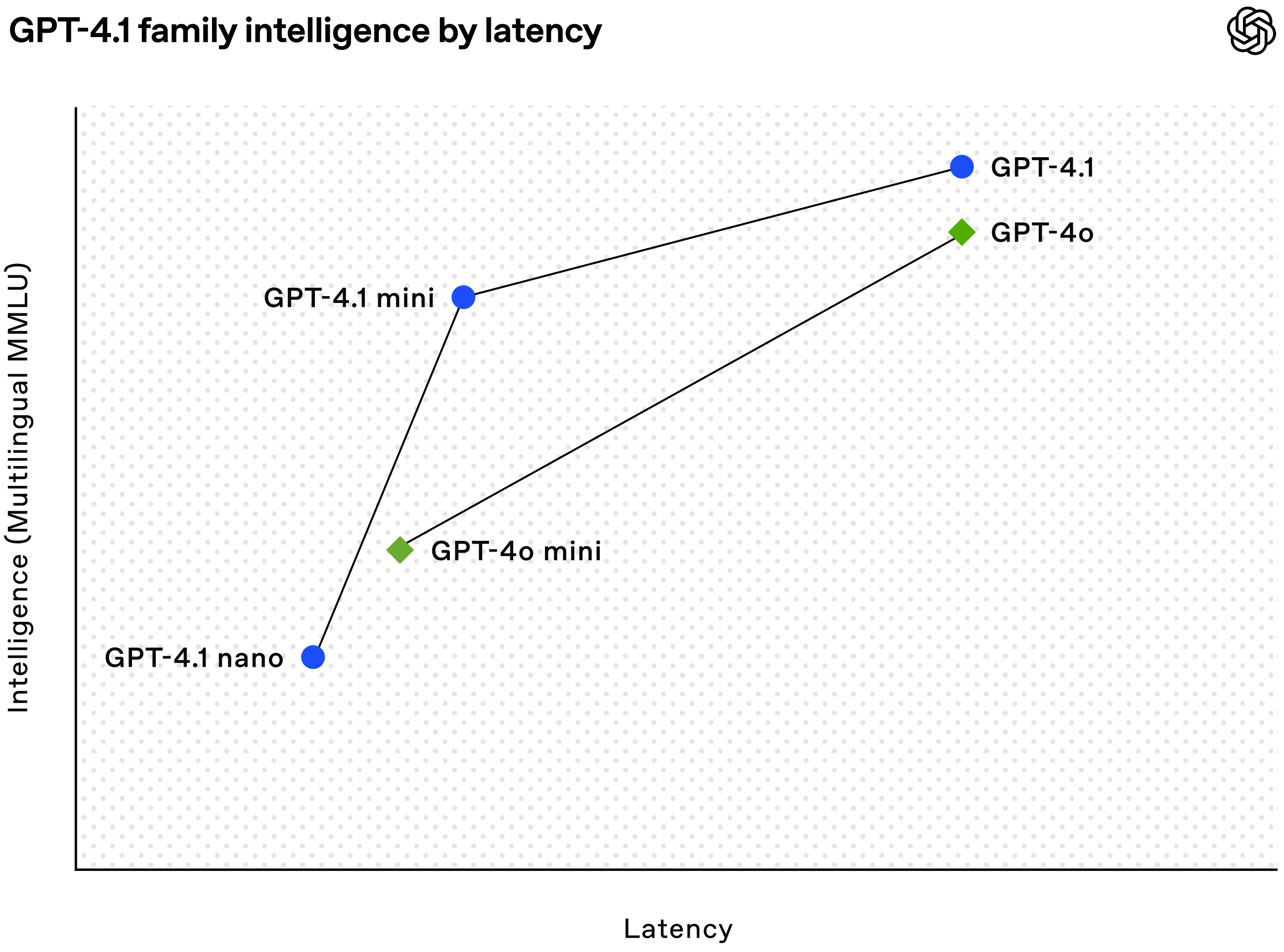
GPT-4.1 mini is a significant leap in small model performance, even beating GPT-4o in many benchmarks. It matches or exceeds GPT-4o in intelligence evals while reducing latency by nearly half and reducing cost by 83%.

For tasks that demand low latency, GPT-4.1 nano is our fastest and cheapest model available. It delivers exceptional performance at a small size with its 1 million token context window, and scores 80.1% on MMLU, 50.3% on GPQA, and 9.8% on Aider

当本 标记提供了宝贵的见解，我们训练这些模型时注重实际应用价值。与开发者的密切合作与伙伴关系

社区使我们能够为其应用中最重要的任务优化这些模型。

为此，GPT-4.1 模型系列在更低的成本下提供了卓越的性能。这些模型在延迟曲线的每个点上都推动性能向前发展。



GPT-4.1 mini 在小模型性能方面实现了重大飞跃，甚至在许多基准测试中超越了 GPT-4o。它在智能评估中与 GPT-4o 相当或优于它，同时将延迟降低了近一半，成本降低了83%。

对于需要低延迟的任务，GPT-4.1 nano 是我们最快且最经济的模型。它凭借其 100 万令牌的上下文窗口，在体积小巧的同时提供卓越的性能，在 MMLU 上得分为 80.1%，在 GPQA 上得分为 50.3%，在 Aider 上得分为 9.8%。

These improvements in instruction following reliability and long context comprehension also make the GPT-4.1 models considerably more effective at powering agents, or systems that can independently accomplish tasks on behalf of users. When combined with primitives like the Responses API, developers can now build agents that are more useful and reliable at real-world software engineering, extracting insights from large documents, resolving customer requests with minimal hand-holding, and other complex tasks.

Note that GPT-4.1 will only be available via the API. In ChatGPT, many of the improvements in instruction following, coding, and intelligence have been gradually incorporated into the latest version of GPT-4o, and we will continue to incorporate more with future releases.

We will also begin deprecating GPT-4.5 Preview in the API, as GPT-4.1 offers improved or similar performance on many key capabilities at much lower cost and latency. GPT-4.5 Preview will be turned off in three months, on July 14, 2025, to allow time for developers to transition. GPT-4.5 was introduced as a research preview to explore and experiment with a large, compute-intensive model, and we've learned a lot from developer feedback. We'll continue to carry forward the creativity, writing quality, humor, and nuance you told us you appreciate in GPT-4.5 into future API models.

Below, we break down how GPT-4.1 performs across several benchmarks, along with examples from alpha testers like Windsurf, Qodo, Hex, Blue J, Thomson Reuters, and Carlyle that showcase how it performs in production on domain-specific tasks.

Coding

GPT-4.1 is significantly better than GPT-4o at a variety of coding tasks, including agentically solving coding tasks, frontend coding, making fewer extraneous edits, following diff formats reliably, ensuring consistent tool usage, and more.

多语种 C 编码——甚至比 GPT-4o mini 还要高。它非常适合分类或自动补全等任务。

这些在指令遵循可靠性和长上下文理解方面的改进，也使得GPT-4.1模型在驱动代理或能够代表用户独立完成任务的系统方面变得更加高效。当结合诸如Responses API之类的基础功能时，开发者现在可以构建在实际软件工程中更有用、更可靠的代理，能够从大量文档中提取见解、以最少的引导解决客户请求，以及完成其他复杂任务。

请注意，GPT-4.1 仅通过 API 提供。在 ChatGPT 中，许多关于指令遵循、编码和智能方面的改进已逐步融入最新版本的 GPT-4o，我们将继续在未来的版本中加入更多改进。

我们也将开始在API中逐步淘汰GPT-4.5预览版，因为GPT-4.1在许多关键能力上提供了更优或相似的性能，且成本和延迟都大大降低。GPT-4.5预览版将于2025年7月14日关闭，三个月后，以便开发者有时间进行过渡。GPT-4.5作为一个研究预览版，旨在探索和试验一个大型、计算密集型的模型，我们从开发者的反馈中学到了很多。我们将继续将您在GPT-4.5中所欣赏的创造力、写作质量、幽默感和细腻之处，融入未来的API模型中。

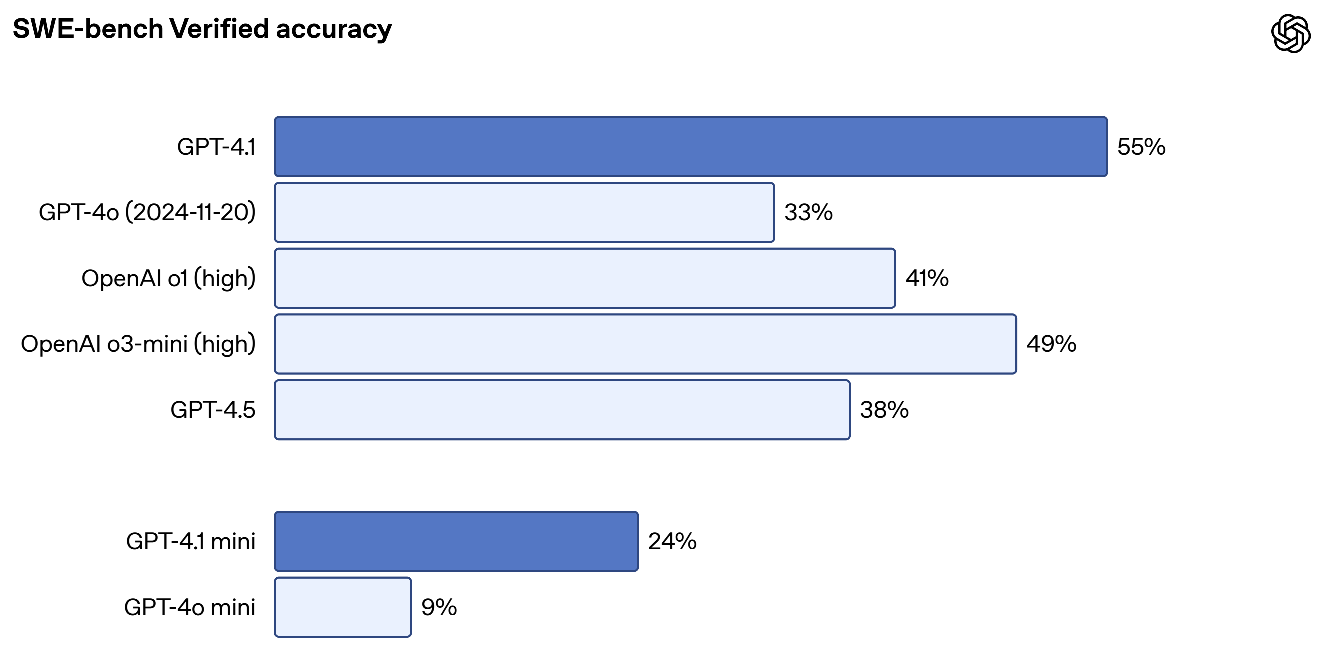
介绍了

下面，我们将详细介绍GPT-4.1在多个基准测试中的表现，以及来自Alpha测试者如Windsurf、Qodo、Hex、Blue J、Thomson Reuters和Carlyle的示例，展示其在实际生产中处理特定领域任务的表现。

编码

GPT-4.1 在多种编码任务中明显优于 GPT-4o，包括自主解决编码任务、前端编码、减少多余的编辑、可靠地遵循差异格式、确保工具使用的一致性等。

improvements in model ability to explore a code repository, finish a task, and produce code that both runs and passes tests.

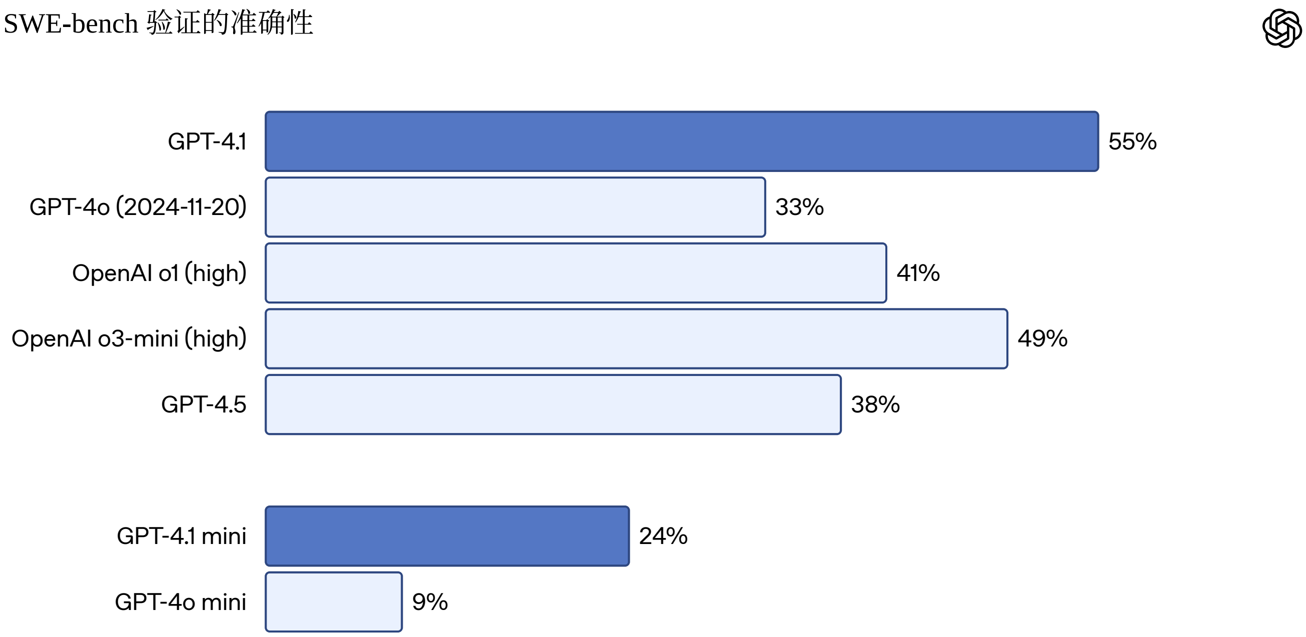


For SWE-bench Verified, a model is given a code repository and issue description, and must generate a patch to solve the issue. Performance is highly dependent on the prompts and tools used. To aid in reproducing and contextualizing our results, we describe our setup for GPT-4.1 here. Our scores omit 23 of 500 problems whose solutions could not run on our infrastructure; if these are conservatively scored as 0, the 54.6% score becomes 52.1%.

For API developers looking to edit large files, GPT-4.1 is much more reliable at code diffs across a range of formats. GPT-4.1 more than doubles GPT-4o's score on Aider's polyglot diff benchmark, and even beats GPT-4.5 by 8%$_{\text{abs.}}$This evaluation is both a measure of coding capabilities across various programming languages and a measure of model ability to produce changes in whole and diff formats. We've specifically trained GPT-4.1 to follow diff formats more reliably, which allows developers to save both cost and latency by only having the model output changed lines, rather than rewriting an entire file. For best code diff performance, please refer to our prompting guide. For developers who prefer rewriting entire files, we've increased output token

关于SWE-bench 已验证，是衡量实际软件工程技能的指标，GPT-4.1 完成了 54.6% 的任务，而 GPT-4o 完成了 33.2%（2024-11-20）。这反映出

在模型能力方面的改进，包括探索代码仓库、完成任务，以及生成既能运行又能通过测试的代码。

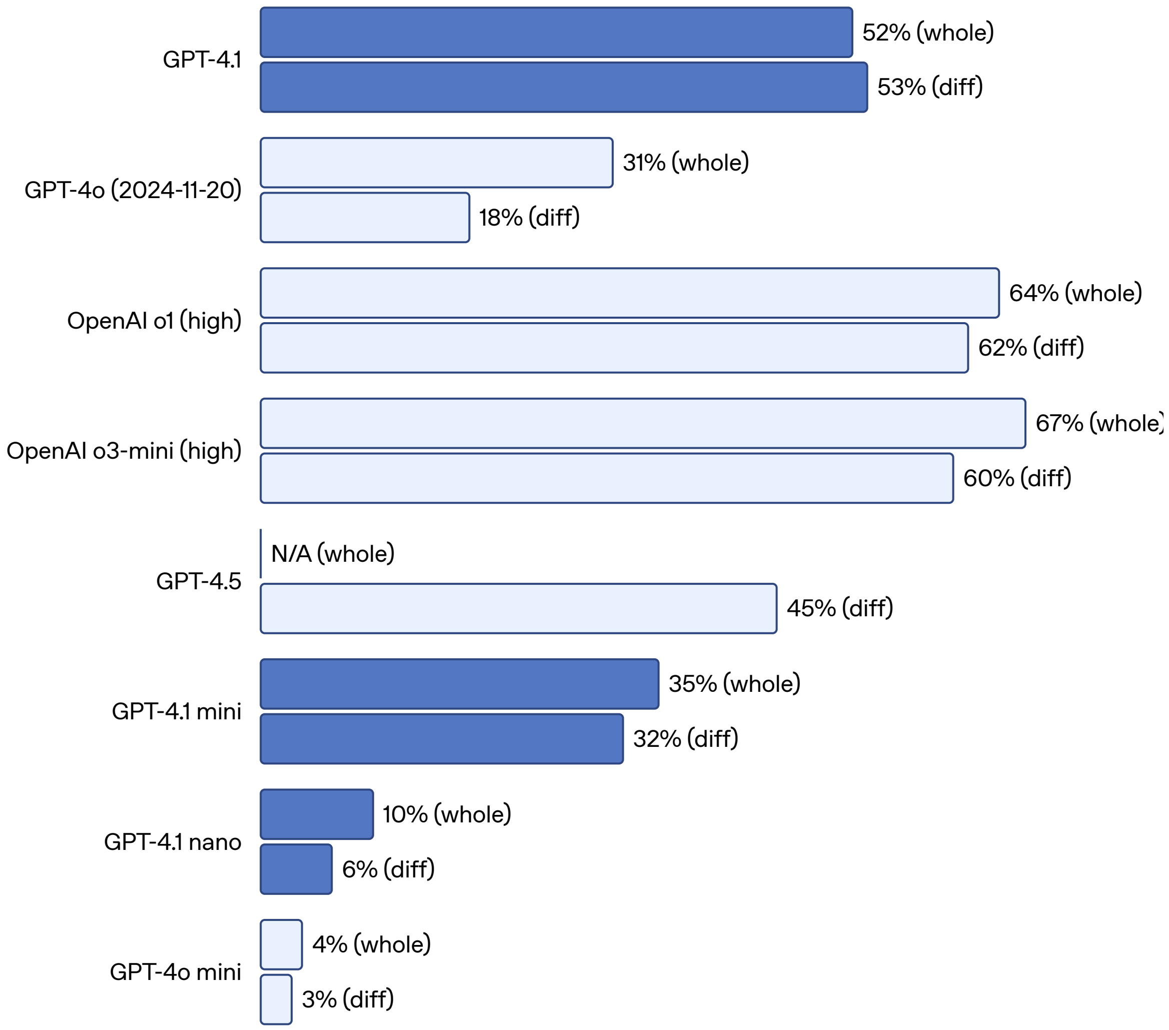


For SWE-bench Verified, a model is given a code repository and issue description, and must generate a patch to solve the issue. Performance is highly dependent on the prompts and tools used. To aid in reproducing and contextualizing our results, we describe our setup for GPT-4.1 here. Our scores omit 23 of 500 problems whose solutions could not run on our infrastructure; if these are conservatively scored as 0, the 54.6% score becomes 52.1%.

对于希望编辑大文件的API开发者来说，GPT-4.1在各种格式的代码差异方面更为可靠。GPT-4.1在Aider的多语言差异基准测试中得分比GPT-4o高出一倍多，甚至比GPT-4.5高出8%。这一评估既衡量了在多种编程语言中的编码能力，也衡量了模型在生成完整和差异格式变更方面的能力。我们特别训练了GPT-4.1以更可靠地遵循差异格式，这使得开发者可以通过只输出变更的行而不是重写整个文件，节省成本和延迟。为了获得最佳的代码差异性能，请参考我们的提示指南。对于偏好重写整个文件的开发者，我们已增加输出令牌数。

绝对值。

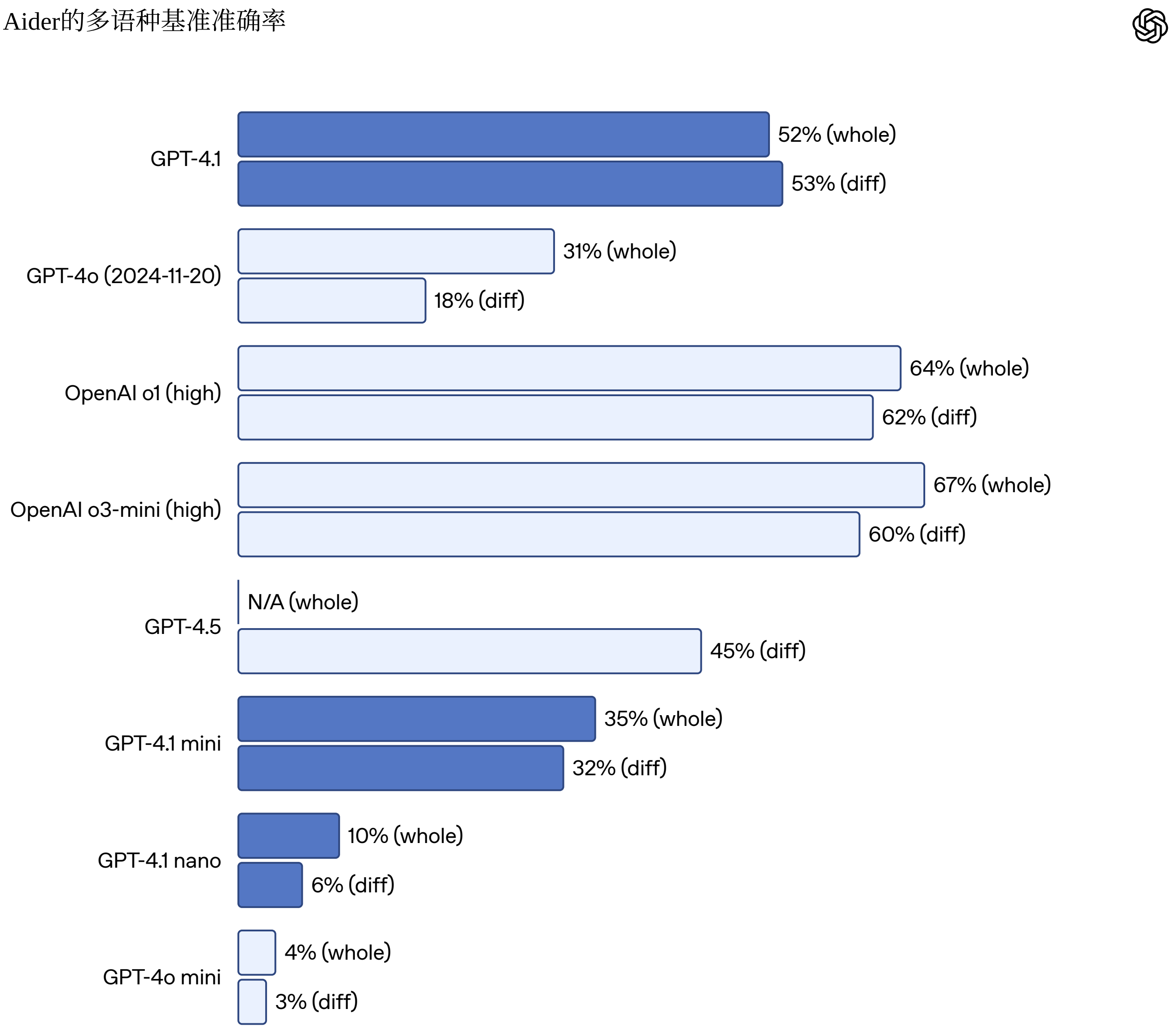
Aider's polyglot benchmark accuracy



In Aider's polyglot benchmark, models solve coding exercises from Exercism by editing source files, with one retry allowed. The 'whole' format requires the model to rewrite the entire file, which can be slow and costly. The 'diff' format requires the model to write a series of search/replace blocks.

GPT-4.1 also substantially improves upon GPT-4o in frontend coding, and is capable of creating web apps that are more functional and aesthetically pleasing. In our head-to-

G的极限 PT-4.1 至 32,768 个标记（高于 GPT-4o 的 16,384 个标记）。我们还建议使用预测输出以减少完整文件重写的延迟。



In Aider's polyglot benchmark, models solve coding exercises from Exercism by editing source files, with one retry allowed. The 'whole' format requires the model to rewrite the entire file, which can be slow and costly. The 'diff' format requires the model to write a series of search/replace blocks.

GPT-4.1 在前端编码方面也大幅优于 GPT-4o，能够创建功能更强大、外观更美观的网页应用。在我们的头对-

Prompt: Make a flashcard web application. The user should be able to create flashcards, search through their existing flashcards, review flashcards, and see statistics on flashcards reviewed. Preload ten cards containing a Hindi word or phrase and its English translation. Review interface: In the review interface, clicking or pressing Space should flip the card with a smooth 3-D animation to reveal the translation. Pressing the arrow keys should navigate through cards. Search interface: The search bar should dynamically provide a list of results as the user types in a query. Statistics interface: The stats page should show a graph of the number of cards the user has reviewed, and the percentage they have gotten correct. Create cards interface: The create cards page should allow the user to specify the front and back of a flashcard and add to the user's collection. Each of these interfaces should be accessible in the sidebar. Generate a single page React app (put all styles inline).

GPT-4o

头部 com 对比，付费人工评分员在80%的情况下更喜欢GPT-4.1的网站而非GPT-4o的网站。

提示：制作一个闪卡网页应用程序。用户应该能够创建闪卡、搜索已有的闪卡、复习闪卡，并查看已复习闪卡的统计数据。预加载十张包含印地语单词或短语及其英文翻译的卡片。复习界面：在复习界面中，点击或按空格键应通过平滑的3D动画翻转卡片，显示翻译。按箭头键应在卡片之间导航。搜索界面：搜索栏应在用户输入查询时动态提供结果列表。统计界面：统计页面应显示用户已复习卡片数量的图表，以及他们答对的百分比。创建卡片界面：创建卡片页面应允许用户指定闪卡的正面和背面，并添加到用户的收藏中。每个界面都应在侧边栏中可访问。生成一个单页React应用（所有样式内联）。

GPT-4o

GPT-4.1

Beyond the benchmarks above, GPT-4.1 is better at following formats more reliably and makes extraneous edits less frequently. In our internal evals, extraneous edits on code dropped from 9% with GPT-4o to 2% with GPT-4.1.

Real world examples

Windsurf: GPT-4.1 scores 60% higher than GPT-4o on Windsurf's internal coding benchmark, which correlates strongly with how often code changes are accepted on the first review. Their users noted that it was 30% more efficient in tool calling and about 50% less likely to repeat unnecessary edits or read code in overly narrow, incremental steps. These improvements translate into faster iteration and smoother workflows for engineering teams.

Qodo: Qodo tested GPT-4.1 head-to-head against other leading models on generating high-quality code reviews from GitHub pull requests using a methodology inspired by their fine-tuning benchmark. Across 200 meaningful real-world pull requests with the

GPT-4.1

除了上述基准测试之外，GPT-4.1在更可靠地遵循格式方面表现更佳，并且更少进行多余的编辑。在我们的内部评估中，代码中的多余编辑从GPT-4o的9%下降到GPT-4.1的2%。

现实世界的例子

风帆冲浪：GPT-4.1 在风帆冲浪的内部编码基准测试中得分比 GPT-4o 高出 60%，该基准测试与代码更改在首次审查中被接受的频率密切相关。用户指出，它在工具调用方面的效率提高了 30%，并且在重复不必要的编辑或以过于狭窄、逐步的方式阅读代码方面的可能性降低了约 50%。这些改进转化为工程团队更快的迭代速度和更流畅的工作流程。

Qodo：Qodo 使用一种受其微调基准启发的方法，在与其他领先模型的正面对抗中测试了 GPT-4.1 在从 GitHub 拉取请求生成高质量代码审查方面的能力。在 200 个有意义的真实世界拉取请求中，

when not to make suggestions) and comprehensiveness (providing thorough analysis when warranted), while maintaining focus on truly critical issues.

Instruction following

GPT-4.1 follows instructions more reliably, and we've measured significant improvements across a variety of instruction following evals.

We developed an internal eval for instruction following to track model performance across a number of dimensions and in several key categories of instruction following, including:

- **Format following.** Providing instructions that specify a custom format for the model's response, such as XML, YAML, Markdown, etc.
- **Negative instructions.** Specifying behavior the model should avoid. (Example: "Don't ask the user to contact support")
- **Ordered instructions.** Providing a set of instructions the model must follow in a given order. (Example: "First ask for the user's name, then ask for their email")
- **Content requirements.** Outputting content that includes certain information. (Example: "Always include amount of protein when writing a nutrition plan")
- **Ranking.** Ordering the output in a particular way. (Example: "Sort the response by population count")
- **Overconfidence.** Instructing the model to say "I don't know" or similar if requested information isn't available, or the request doesn't fall in a given category. (Example: "If you do not know the answer, provide the support contact email")

These categories are the result of feedback from developers regarding which facets of instruction following are most relevant and important to them. Within each category,

相同的促销pts 和条件，他们发现 GPT-4.1 在 55% 的情况下提供了更好的建议。值得注意的是，他们发现 GPT-4.1 在精确度（知道 {v*}）方面都表现出色。

何时不提出建议）以及全面性（在必要时提供全面分析），同时保持对真正关键问题的关注。

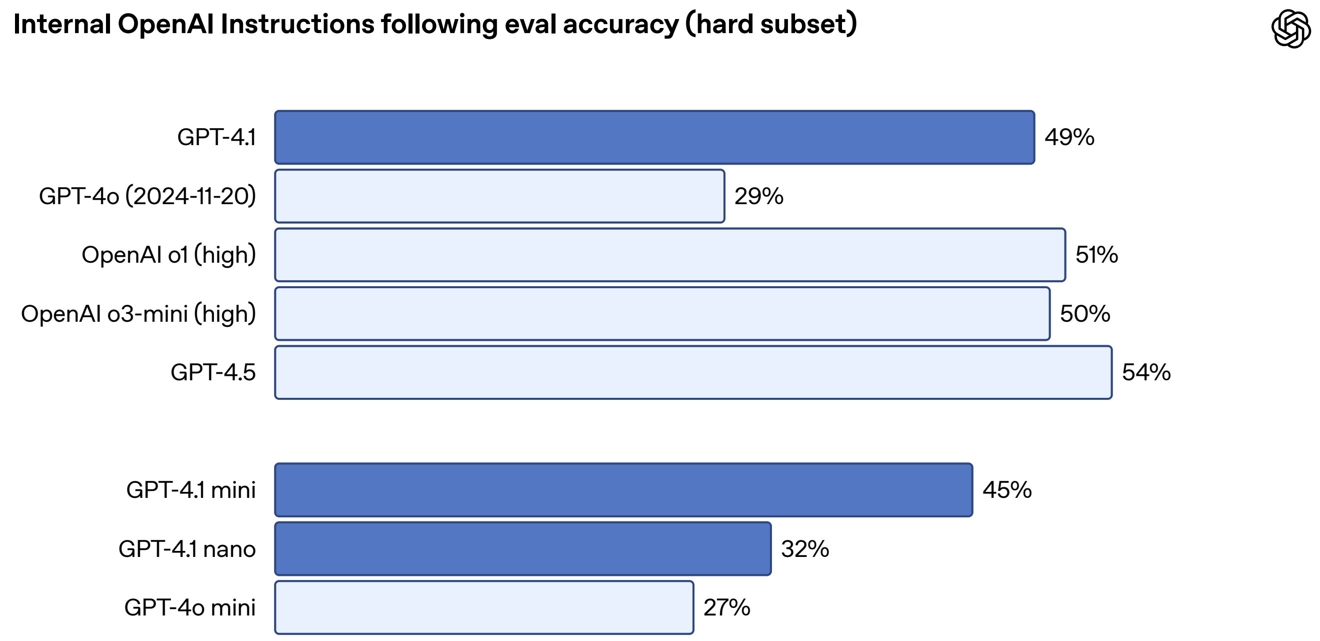
遵循指令

GPT-4.1 更加可靠地遵循指令，我们在各种指令执行评估中都测量到了显著的改进。

我们开发了一个内部评估，用于跟踪模型在多个维度和几个关键指令执行类别中的表现，包括：

- 格式如下。提供指示，指定模型响应的自定义格式，例如 XML、YAML、Markdown 等。
- 负面指令。指定模型应避免的行为。（示例：“不要让用户联系支持”）
- 有序指令。提供一组模型必须按照给定顺序遵循的指令。（示例：“首先询问用户的姓名，然后询问他们的电子邮件”）
- 内容要求。输出包含特定信息的内容。（例如：“在制定营养计划时，始终包括蛋白质的含量”）
- 排名。以特定方式对输出进行排序。（示例：“按人口数量排序响应”）
- 过度自信。当指示模型在请求信息不可用或请求不属于给定类别时，说“我不知道”或类似的话。（示例：“如果你不知道答案，请提供支持联系邮箱”）

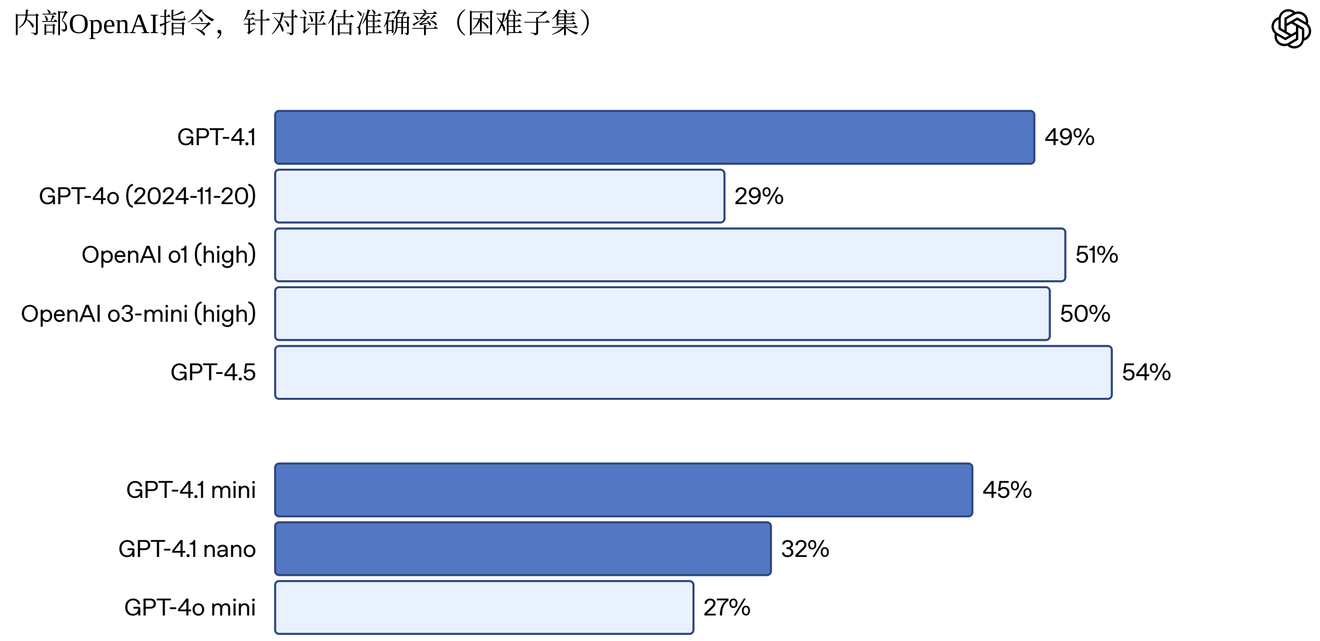
这些类别是开发者关于指令执行中哪些方面最相关和重要的反馈的结果。在每个类别中，



Our internal instruction following eval is based on real developer use cases and feedback, covering tasks of varying complexity coupled with instructions on formatting, verbosity, length, and more.

Multi-turn instruction following is critical for many developers—it's important for the model to maintain coherence deep into a conversation, and keep track of what the user told it earlier. We've trained GPT-4.1 to be better able to pick out information from past messages in the conversation, allowing for more natural conversations. The MultiChallenge benchmark from Scale is a useful measure of this capability, and GPT-4.1 performs 10.5%$_{abs}$ better than GPT-4o.

我们已经分为轻松、中等和困难的提示。GPT-4.1 在特别是困难提示方面比 GPT-4o 有显著的提升。

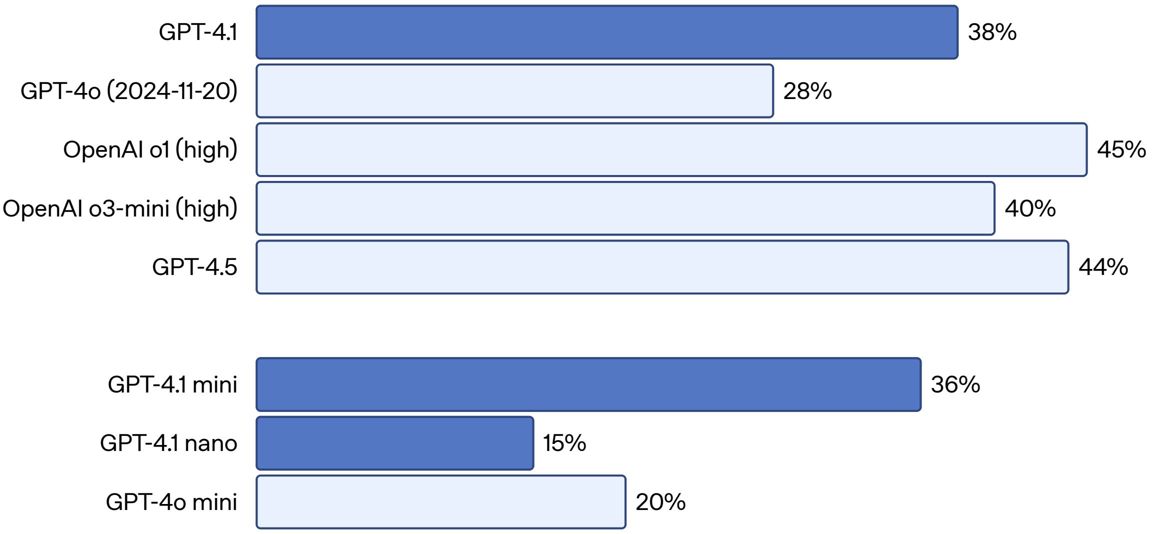


内部OpenAI指令，针对评估准确率（困难子集）

Our internal instruction following eval is based on real developer use cases and feedback, covering tasks of varying complexity coupled with instructions on formatting, verbosity, length, and more.

多轮指令跟随对许多开发者来说至关重要——模型在对话中保持连贯性并跟踪用户早期提供的信息非常重要。我们已经训练了 GPT-4.1，使其更好地从对话的过去消息中提取信息，从而实现更自然的对话。Scale 的 MultiChallenge 基准测试是衡量这一能力的有用指标，而 GPT-4.1 的表现比 GPT-4o 高出 10.5%。

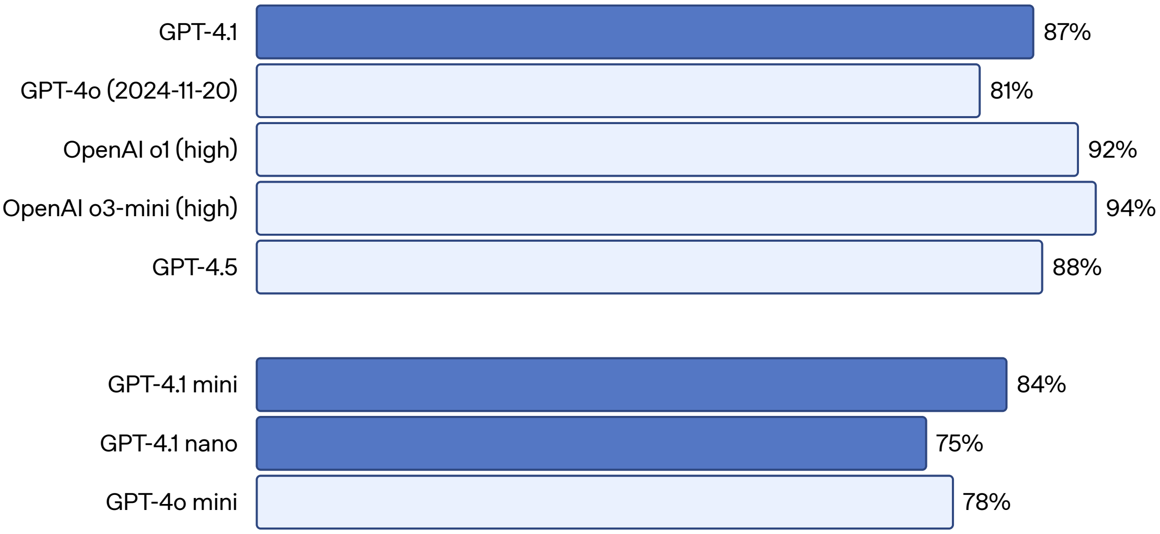
绝对值



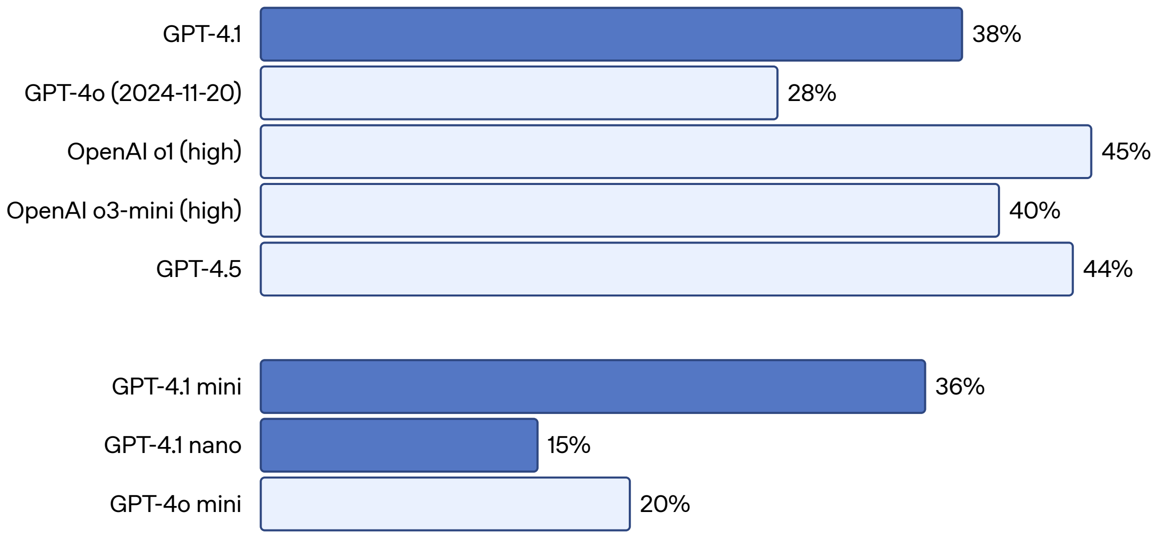
In MultiChallenge, models are challenged on multi-turn conversations to properly use four types of information from previous messages.

GPT-4.1 also scores 87.4% on IFEval, compared to 81.0% for GPT-4o. IFEval uses prompts with verifiable instructions (for example, specifying content length or avoiding certain terms or formats).

IFEval accuracy



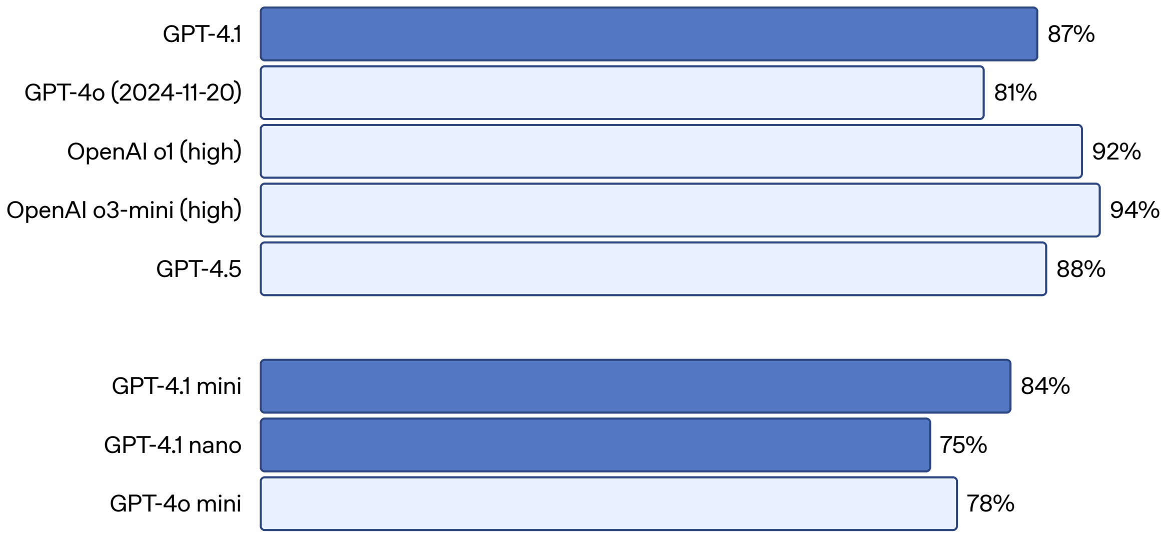
MultiChall 工程精度



In MultiChallenge, models are challenged on multi-turn conversations to properly use four types of information from previous messages.

GPT-4.1 在 IFEval 上的得分也达到了87.4%，而 GPT-4o 为81.0%。IFEval 使用带有可验证指令的提示（例如，指定内容长度或避免某些术语或格式）。

IFEval 准确率



Better instruction following makes existing applications more reliable, and enables new applications previously limited by poor reliability. Early testers noted that GPT-4.1 can be more literal, so we recommend being explicit and specific in prompts. For more on prompting best practices for GPT-4.1, please refer to the prompting guide.

Real world examples

Blue J: GPT-4.1 was 53% more accurate than GPT-4o on an internal benchmark of Blue J's most challenging real-world tax scenarios. This jump in accuracy—key to both system performance and user satisfaction—highlights GPT-4.1's improved comprehension of complex regulations and its ability to follow nuanced instructions over long contexts. For Blue J users, that means faster, more reliable tax research and more time for high-value advisory work.

Hex: GPT-4.1 delivered a nearly 2× improvement on Hex's most challenging SQL evaluation set, showcasing significant gains in instruction following and semantic understanding. The model was more reliable in selecting the correct tables from large, ambiguous schemas—an upstream decision point that directly impacts overall accuracy and is difficult to tune through prompting alone. For Hex, this resulted in a measurable reduction in manual debugging and a faster path to production-grade workflows.

Long Context

GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano can process up to 1 million tokens of context—up from 128,000 for previous GPT-4o models. 1 million tokens is more than 8 copies of the entire React codebase, so long context is a great fit for processing large codebases, or lots of long documents.

In IFEval, models must generate answers that comply with various instructions.

更好的指令遵循使现有应用更可靠，并使以前因可靠性差而受限的新应用成为可能。早期测试者指出，GPT-4.1 可以更字面，因此我们建议在提示中明确和具体。关于 GPT-4.1 的提示最佳实践，更多信息请参阅提示指南。

现实世界的例子

Blue J：GPT-4.1在Blue J最具挑战性的实际税务场景内部基准测试中，比GPT-4o的准确率高出53%。这一准确率的提升——对系统性能和用户满意度都至关重要——突显了GPT-4.1在理解复杂法规方面的改进，以及其在长篇上下文中遵循细微指令的能力。对于Blue J的用户来说，这意味着更快、更可靠的税务研究，以及更多时间用于高价值的咨询工作。

Hex： GPT-4.1 在 Hex 最具挑战性的 SQL 评估集上实现了近 2× 的提升，展现出在指令执行和语义理解方面的显著进步。该模型在从大型模糊架构中选择正确的表格方面更为可靠——这是一个直接影响整体准确性的上游决策点，仅通过提示调优难以实现。对于 Hex 来说，这带来了可衡量的手动调试减少和更快的生产级工作流程的路径。

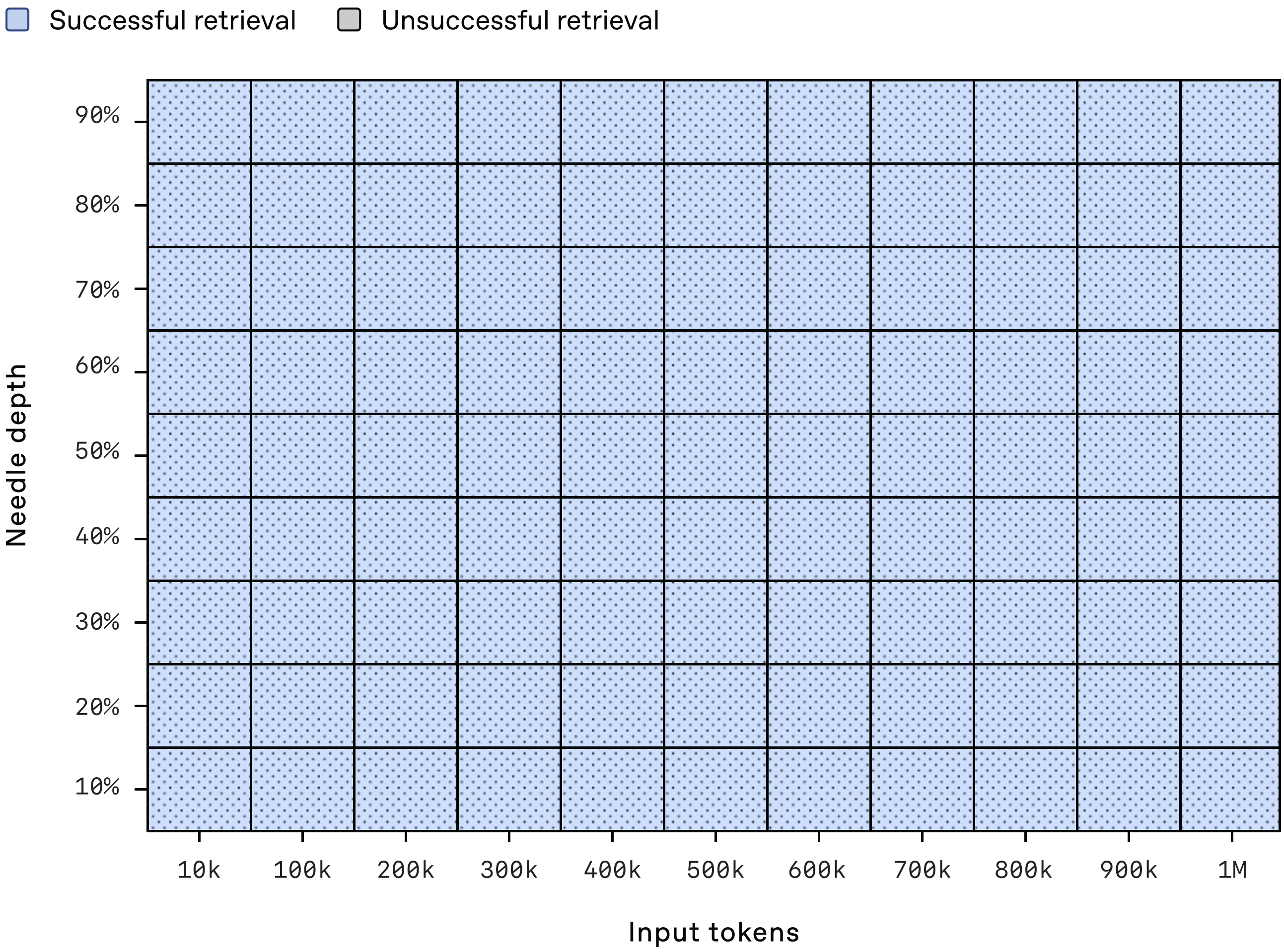
长上下文

GPT-4.1、GPT-4.1 mini 和 GPT-4.1 nano 可以处理多达 100 万个标记的上下文——相比之前的 GPT-4o 模型的 128,000 个标记有了显著提升。100 万个标记大约相当于整个 React 代码库的 8 份以上，因此长上下文非常适合处理大型代码库或大量长文档。

text, and ignoring distractors across long and short context lengths. Long-context understanding is a critical capability for applications across legal, coding, customer support, and many other domains.

Below, we demonstrate GPT-4.1's ability to retrieve a small hidden piece of information (a "needle") positioned at various points within the context window. GPT-4.1 consistently retrieves the needle accurately at all positions and all context lengths, all the way up to 1 million tokens. It is effectively able to pull out relevant details for the task at hand regardless of their position in the input.

GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano needle in a haystack accuracy

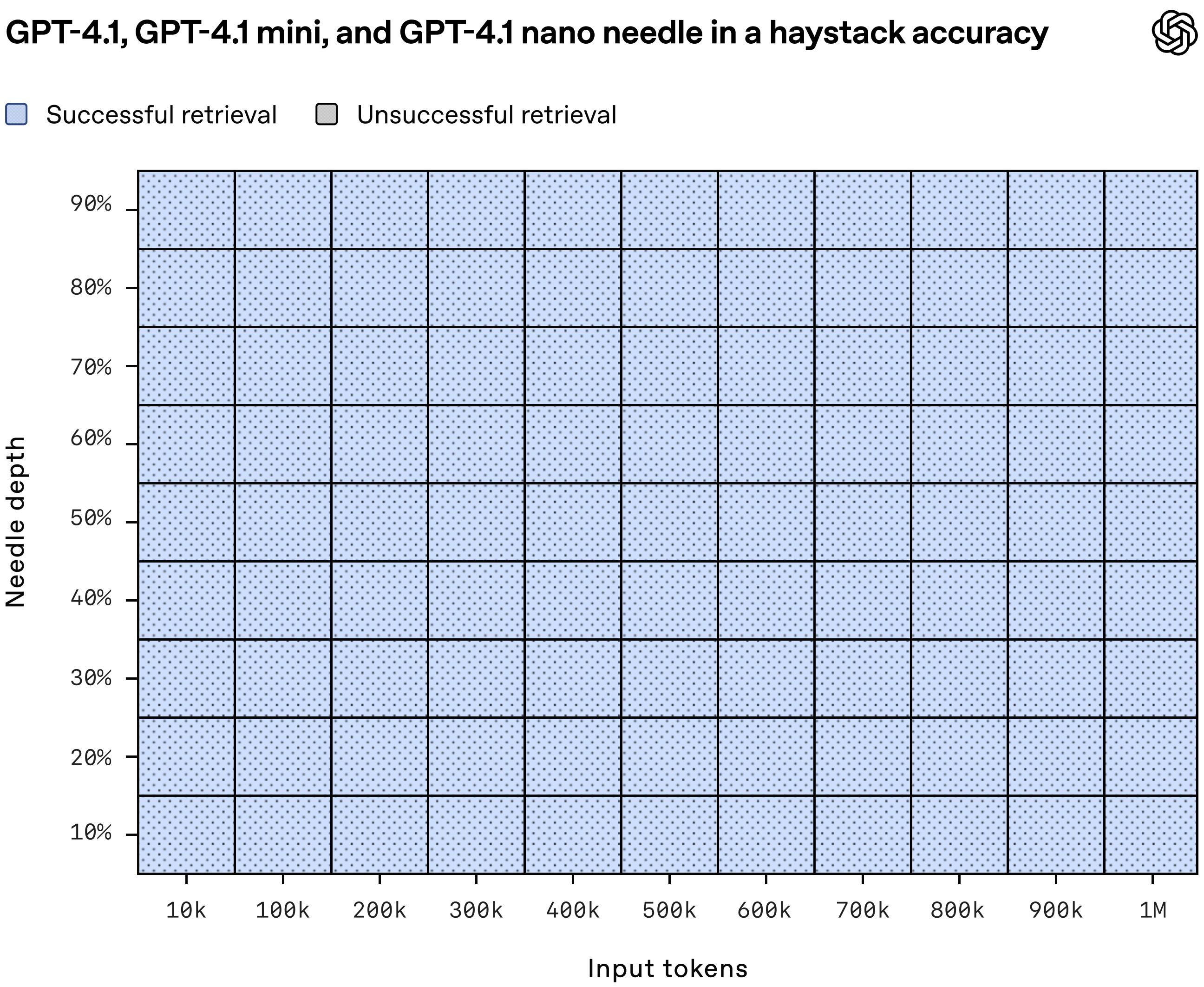


In our internal needle in a haystack eval, GPT-4.1, GPT-4.1 mini, and GPT 4.1 nano are all able to retrieve the needle at all positions in the context up to 1M.

我们训练 d GPT-4.1 能够可靠地关注整个 1 百万长度的上下文信息。我们还训练它比 GPT-4o 更加可靠，能够更好地注意到相关内容。

文本，并忽略跨越长短上下文长度的干扰项。长上下文理解是法律、编码、客户支持以及许多其他领域应用中的关键能力。

以下，我们展示GPT-4.1检索隐藏信息片段（“针”）的能力，该片段位于上下文窗口的不同位置。GPT-4.1在所有位置和所有上下文长度下都能准确检索出“针”，一直到100万标记。它实际上能够提取与任务相关的细节，无论它们在输入中的位置如何。



In our internal needle in a haystack eval, GPT-4.1, GPT-4.1 mini, and GPT 4.1 nano are all able to retrieve the needle at all positions in the context up to 1M.

multiple pieces of information, and to understand those pieces in relation to each other. To showcase this capability, we’re open-sourcing a new eval: OpenAI-MRCR (Multi-Round Coreference).

OpenAI-MRCR tests the model’s ability to find and disambiguate between multiple needles well hidden in context. The evaluation consists of multi-turn synthetic conversations between a user and assistant where the user asks for a piece of writing about a topic, for example "write a poem about tapirs" or "write a blog post about rocks". We then insert two, four, or eight identical requests throughout the context. The model must then retrieve the response corresponding to a specific instance (e.g., “give me the third poem about tapirs”).

The challenge arises from the similarity between these requests and the rest of the context—models can easily be misled by subtle differences, such as a short story about tapirs rather than a poem, or a poem about frogs instead of tapirs. We find that GPT-4.1 outperforms GPT-4o at context lengths up to 128K tokens and maintains strong performance even up to 1 million tokens.

But the task remains hard—even for advanced reasoning models. We’re sharing the eval dataset to encourage further work on real-world long-context retrieval.

2 needle 4 needle 8 needle

然而，很少有实际任务像检索一个单一、明显的针状答案那么直接。我们发现用户通常需要我们的模型进行检索和理解

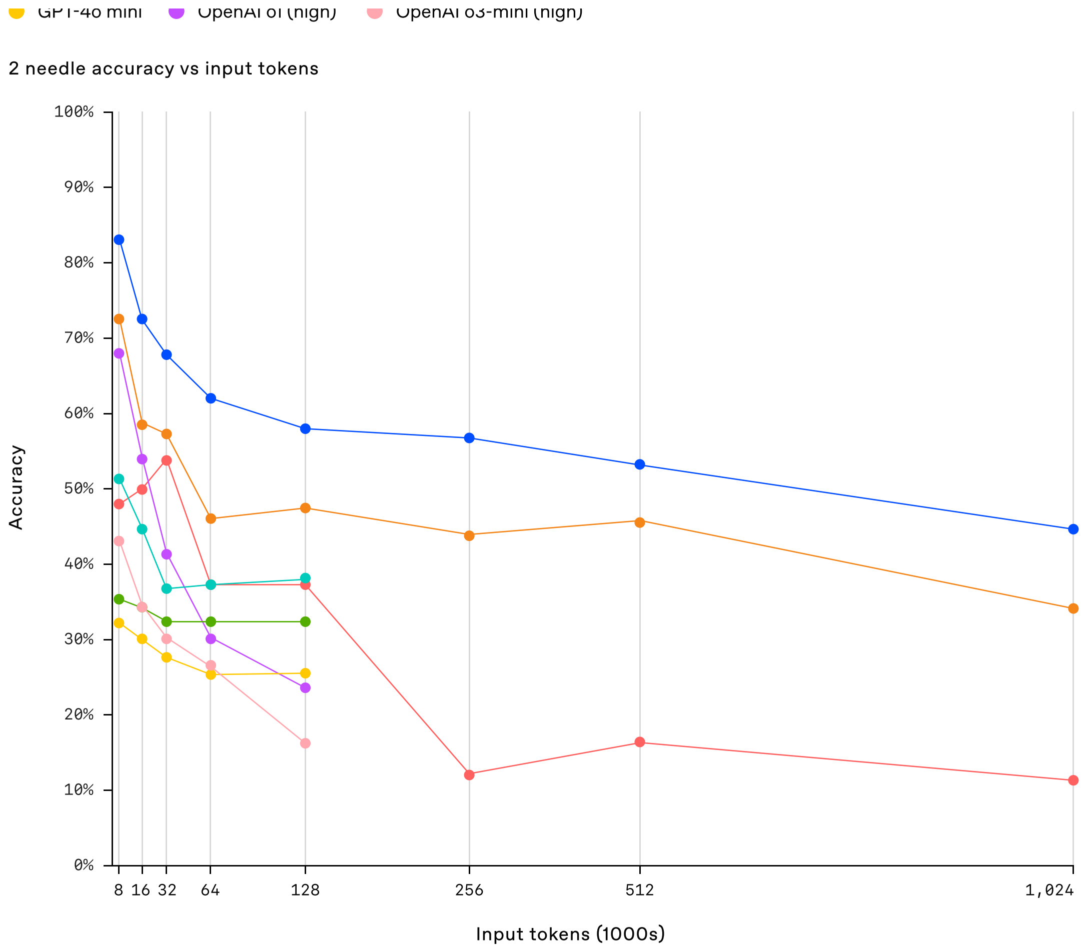
多个信息片段，并理解这些片段之间的关系。为了展示这一能力，我们开源了一个新的评估工具：OpenAI-MRCR（多轮指代消解）。

OpenAI-MRCR 测试模型在上下文中隐藏的多个针之间进行查找和消歧的能力。评估包括用户与助手之间的多轮合成对话，用户会请求关于某个主题的文章，例如“写一首关于貘的诗”或“写一篇关于岩石的博客”。然后，我们在上下文中插入两个、四个或八个相同的请求。模型必须检索出对应特定实例的响应（例如，“给我第三首关于貘的诗”）。

挑战在于这些请求与其余上下文之间的相似性——模型很容易被微妙的差异所误导，比如一篇关于貘的短篇故事而不是一首诗，或者一首关于青蛙而不是貘的诗。我们发现，GPT-4.1在上下文长度达到128K个标记时优于GPT-4o，并且即使在高达1百万个标记的情况下也能保持强劲的性能。

但这个任务仍然很困难——即使对于先进的推理模型也是如此。我们分享这个评估数据集，以鼓励在真实世界长上下文检索方面的进一步研究。

2 needle 4 needle 8 needle



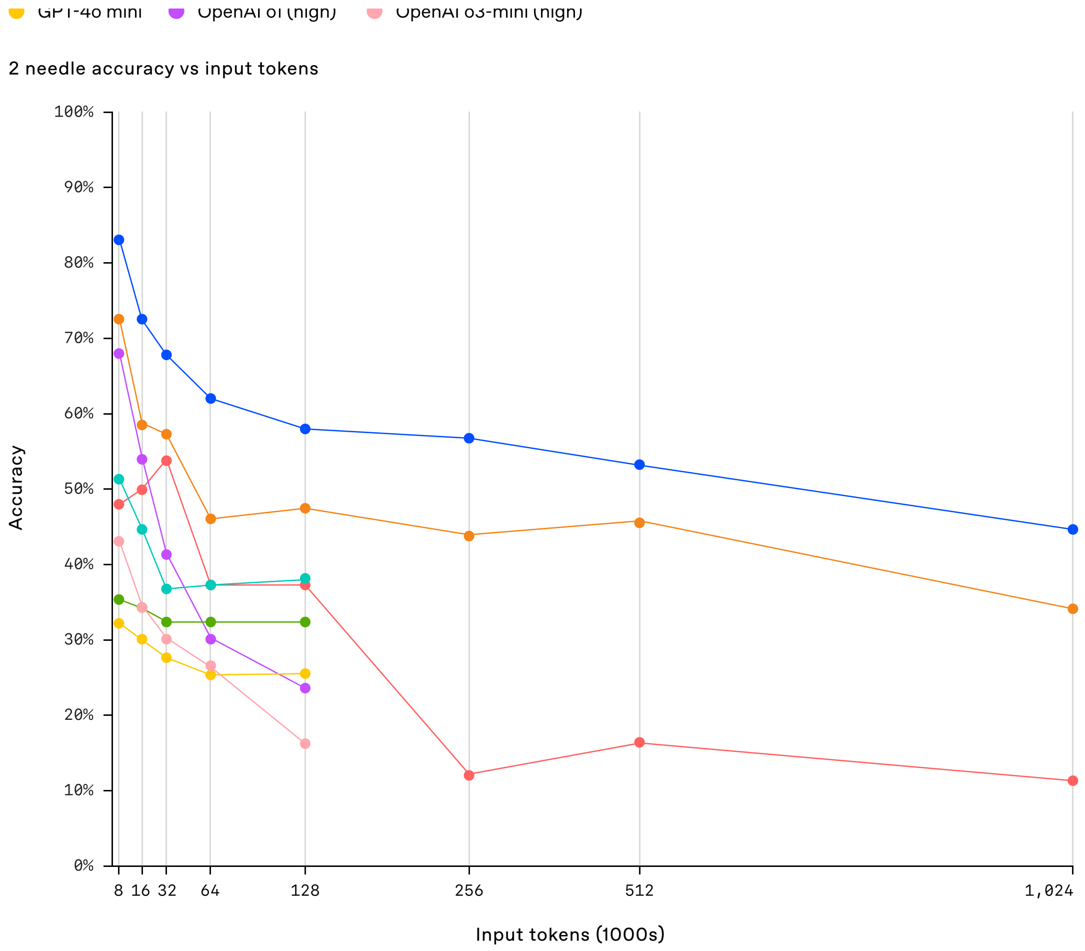
In OpenAI-MRCR, the model must answer a question that involves disambiguating between 2, 4, or 8 user prompts scattered amongst distractors.

Show more

We're also releasing Graphwalks, a dataset for evaluating multi-hop long-context reasoning. Many developer use cases for long context require multiple logical hops within the context, like jumping between multiple files when writing code or cross referencing documents when answering complicated legal questions.

A model (or even a human) could theoretically solve an OpenAI-MRCR problem by doing one pass or read-through of the prompt, but Graphwalks is designed to require reasoning across multiple positions in the context and cannot be solved sequentially.

Graphwalks fills the context window with a directed graph composed of hexadecimal hashes, and then asks the model to perform a breadth-first search (BFS) starting from a random node in the graph. We then ask it to return all nodes at a certain depth.



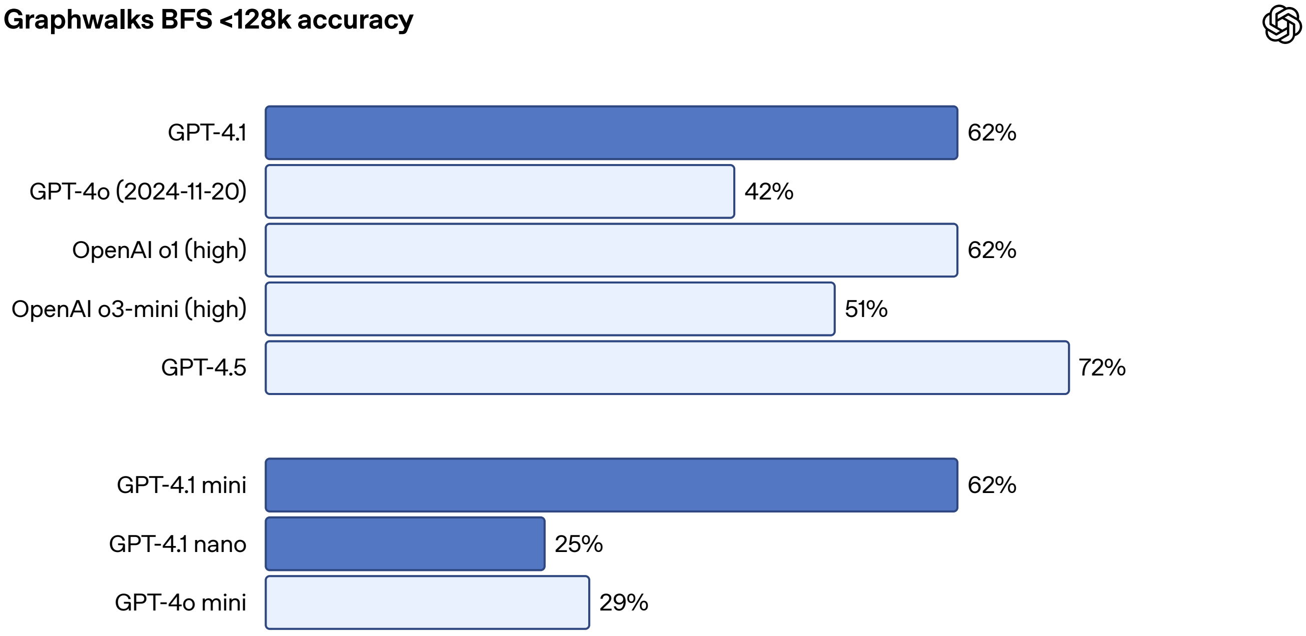
In OpenAI-MRCR, the model must answer a question that involves disambiguating between 2, 4, or 8 user prompts scattered amongst distractors.

显示更多

我们还发布了 Graphwalks，这是一个用于评估多跳长上下文推理的数据集。许多开发者在长上下文中的使用场景需要多次逻辑跳跃，比如在编写代码时在多个文件之间跳转，或在回答复杂法律问题时交叉引用文档。

一个模型（甚至是人类）理论上可以通过对提示进行一次遍历或阅读来解决OpenAI-MRCR问题，但Graphwalks旨在要求跨越上下文中的多个位置进行推理，不能逐步解决。

Graphwalks 用由十六进制哈希组成的有向图填充上下文窗口，然后要求模型从图中的随机节点开始执行广度优先搜索（BFS）。接着，我们让它返回所有在某一深度的节点。



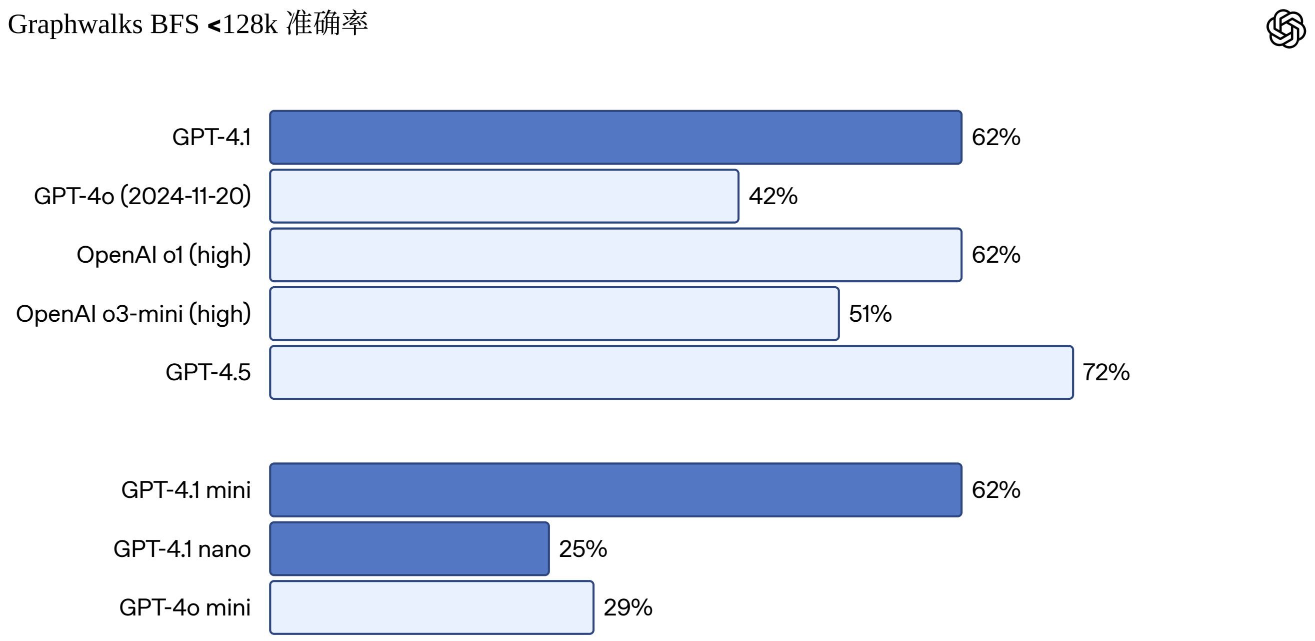
In Graphwalks, a model is asked to perform a breadth-first search from a random node in a large graph.

Benchmarks don’t tell the full story, so we worked with alpha partners to test the performance of GPT-4.1 on their real-world long context tasks.

Real world examples

Thomson Reuters: Thomson Reuters tested GPT-4.1 with CoCounsel, their professional grade AI assistant for legal work. Compared to GPT-4o, they were able to improve multi-document review accuracy by 17% when using GPT-4.1 across internal long-context benchmarks—an essential measure of CoCounsel’s ability to handle complex legal workflows involving multiple, lengthy documents. In particular, they found the model to be highly reliable at maintaining context across sources and accurately

GPT-4.1 ac在这个基准测试中，hieves 达到了 61.7%的准确率，匹配了 o1 的表现，并轻松超越了 GPT-4o。



In Graphwalks, a model is asked to perform a breadth-first search from a random node in a large graph.

基准测试并不能完全反映全部情况，因此我们与alpha合作伙伴合作，测试了GPT-4.1在他们的实际长上下文任务中的表现。

现实世界的例子

汤森路透：汤森路透使用 CoCounsel 测试了 GPT-4.1，这是他们用于法律工作的专业级 AI 助手。与 GPT-4o 相比，他们在内部长上下文基准测试中使用 GPT-4.1 时，能够将多文档审查的准确率提高 17%——这是衡量 CoCounsel 处理涉及多个长篇文档的复杂法律工作流程能力的一个关键指标。特别是，他们发现该模型在保持不同来源的上下文一致性和准确性方面非常可靠。

Carlyle: Carlyle used GPT-4.1 to accurately extract granular financial data across multiple, lengthy documents—including PDFs, Excel files, and other complex formats. Based on their internal evaluations, it performed 50% better on retrieval from very large documents with dense data and was the first model to successfully overcome key limitations seen with other available models, including needle-in-the-haystack retrieval, lost-in-the-middle errors, and multi-hop reasoning across documents.

In addition to model performance and accuracy, developers also need models that respond quickly to keep up with and meet users' needs. We've improved our inference stack to reduce the time to first token, and with prompt caching, you can cut latency even further while saving on costs. In our initial testing, the p95 latency to first token for GPT-4.1 is approximately fifteen seconds with 128,000 tokens of context, and up to half a minute for a million tokens of context. GPT-4.1 mini and nano are faster, e.g., GPT-4.1 nano most often returns the first token in less than five seconds for queries with 128,000 input tokens.

Vision

The GPT-4.1 family is exceptionally strong at image understanding, with GPT-4.1 mini in particular representing a significant leap forward, often beating GPT-4o on image benchmarks.

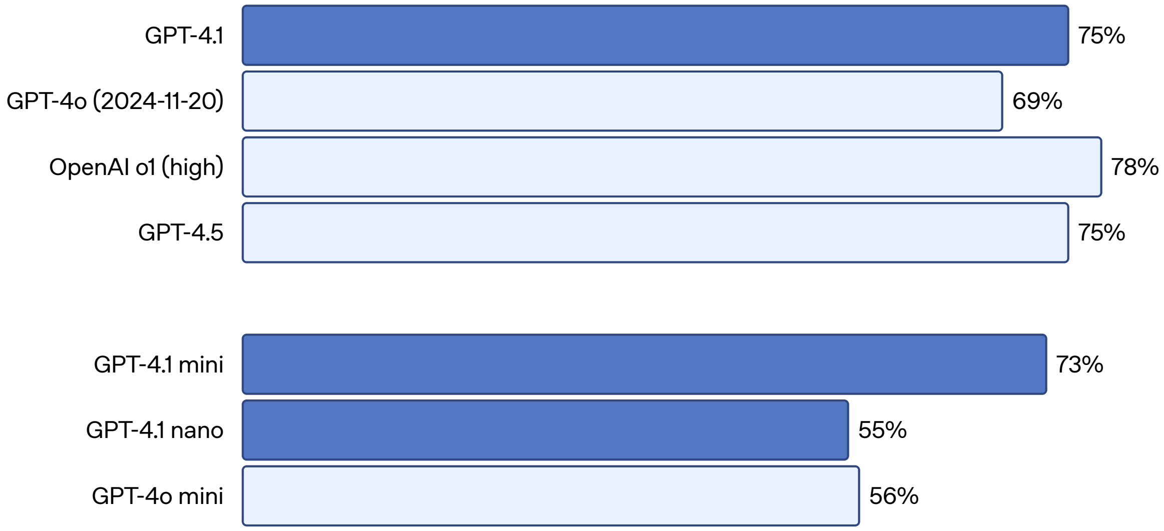
识别细微的文档关系，例如冲突的条款或额外的补充上下文——这些任务对于法律分析和决策至关重要。

卡莱尔：卡莱尔使用GPT-4.1准确提取了多个长篇文档中的细粒度财务数据——包括PDF、Excel文件和其他复杂格式。根据他们的内部评估，在从数据密集的超大文档中检索信息方面，其表现比其他模型高出50%，并且是第一个成功克服其他可用模型所面临的关键限制的模型，包括“大海捞针”式的检索、中途丢失错误以及跨文档的多跳推理。

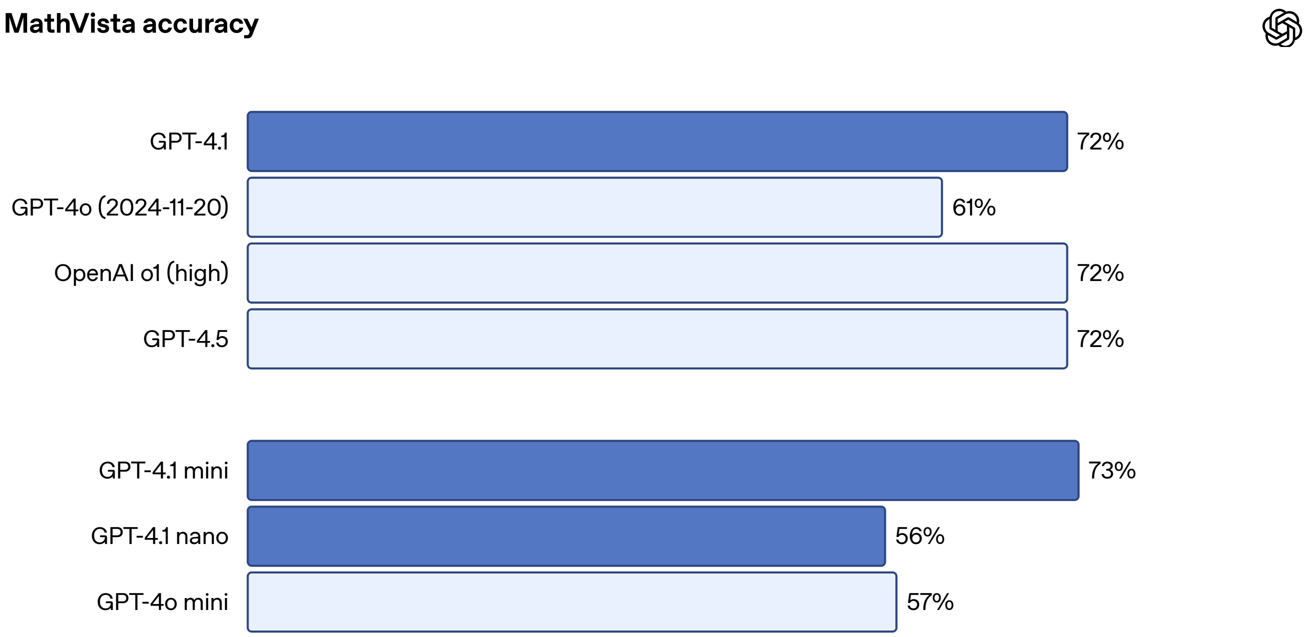
除了模型的性能和准确性之外，开发者还需要模型能够快速响应，以跟上并满足用户的需求。我们已经改进了推理堆栈，缩短了首次生成的时间，并通过提示缓存，您可以进一步降低延迟，同时节省成本。在我们的初步测试中，GPT-4.1 在具有 128,000 个上下文令牌时，首次生成的 p95 延迟大约为十五秒，而在拥有一百万个上下文令牌时，延迟最多可达半分钟。GPT-4.1 mini 和 nano 更快，例如，GPT-4.1 nano 在处理包含 128,000 个输入令牌的查询时，最常在不到五秒的时间内返回第一个令牌。

视觉

GPT-4.1 系列在图像理解方面表现尤为出色，尤其是 GPT-4.1 mini，代表了一个重大飞跃，常常在图像基准测试中超越 GPT-4o。

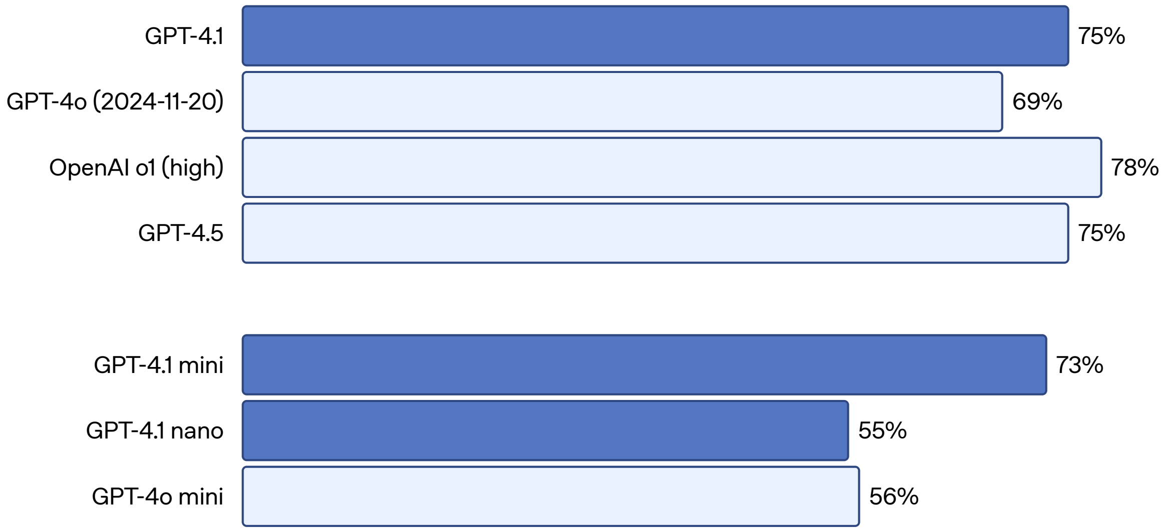


In MMMU, a model answers questions containing charts, diagrams, maps, etc. (Note: even when the image is not included, many answers can still be inferred or guessed from context.)



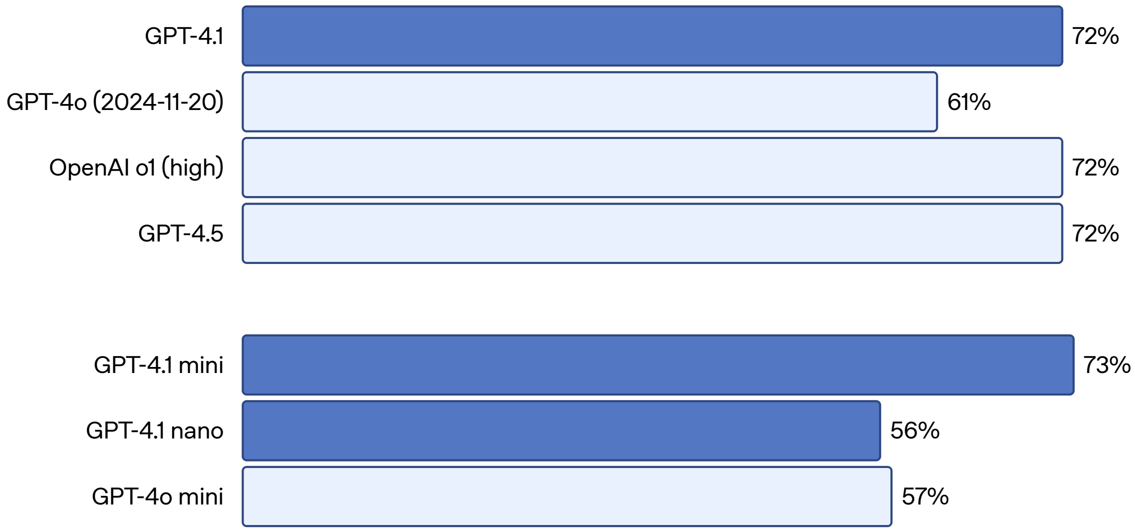
In MathVista, a model solves visual mathematical tasks.

MMMU ac 准确性

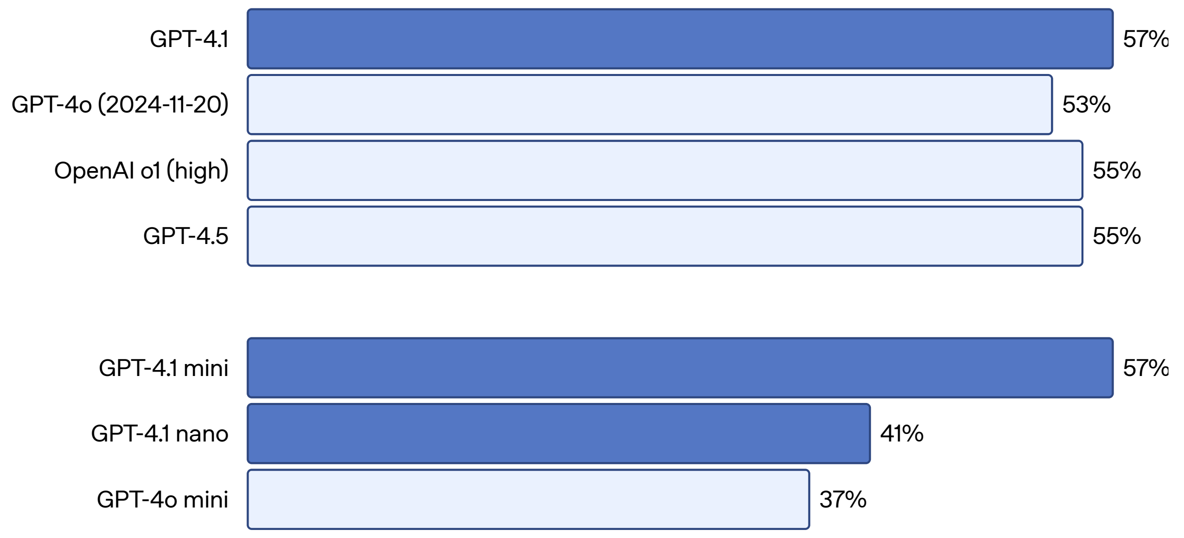


In MMMU, a model answers questions containing charts, diagrams, maps, etc. (Note: even when the image is not included, many answers can still be inferred or guessed from context.)

MathVista 准确率

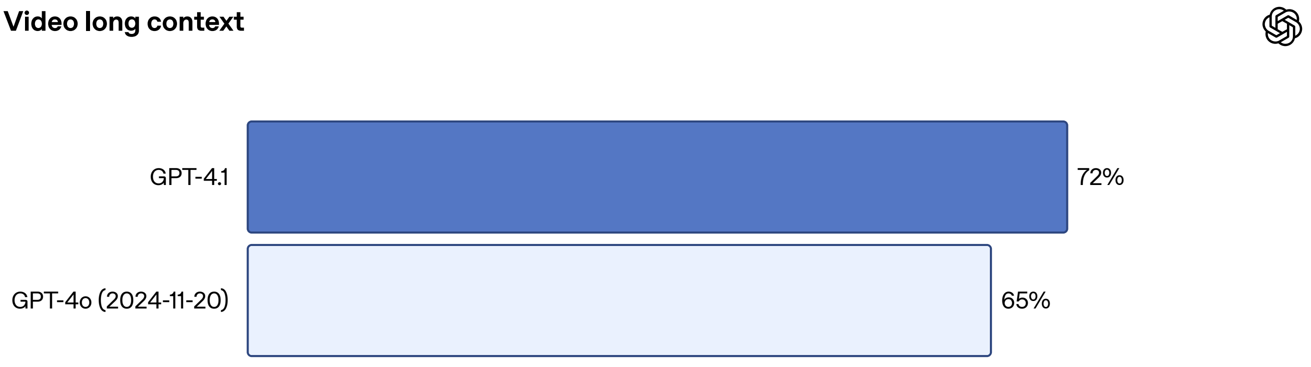


In MathVista, a model solves visual mathematical tasks.



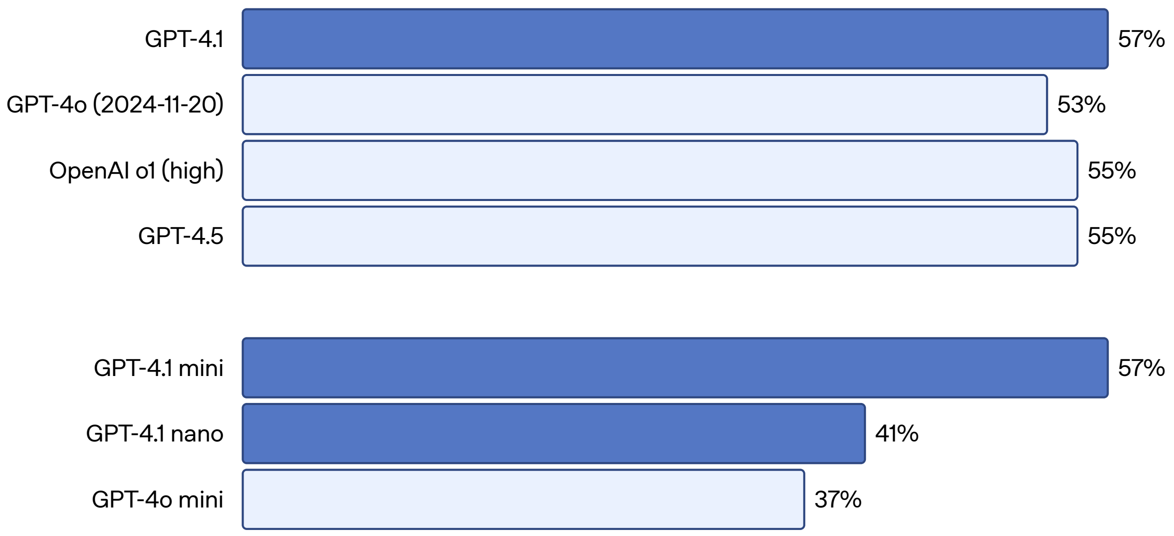
In CharXiv-Reasoning, a model answers questions about charts from scientific papers.

Long context performance is also important for multimodal use cases, such as processing long videos. In Video-MME (long w/o subs), a model answers multiple choice questions based on 30-60 minute long videos with no subtitles. GPT-4.1 achieves state-of-the-art performance, scoring 72.0%, up from 65.3% for GPT-4o.



In Video-MME, a model answers multiple choice questions based on 30-60 minute long videos with no subtitles.

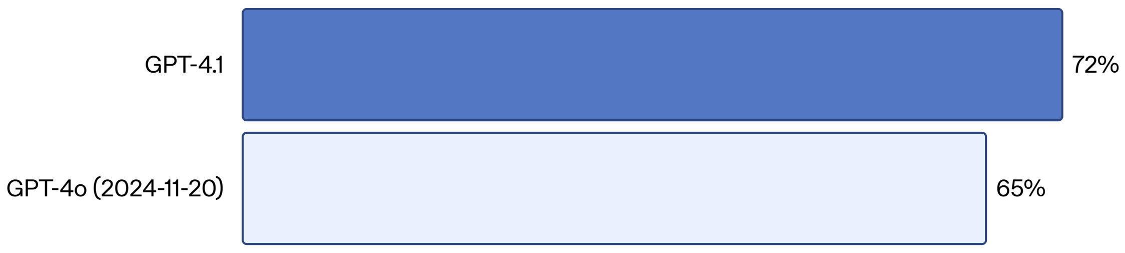
CharXiv-Re 推理准确性



In CharXiv-Reasoning, a model answers questions about charts from scientific papers.

长上下文性能对于多模态用例也很重要，例如处理长视频。在 Video-MME（长无字幕）中，一个模型根据没有字幕的30-60分钟长视频回答多项选择题。GPT-4.1 实现了最先进的性能，得分为72.0%，高于GPT-4o的65.3%。

视频长上下文



In Video-MME, a model answers multiple choice questions based on 30-60 minute long videos with no subtitles.

GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano are available now to all developers.

Through efficiency improvements to our inference systems, we've been able to offer lower prices on the GPT-4.1 series.GPT-4.1 is 26% less expensive than GPT-4o for median queries, and GPT-4.1 nano is our cheapest and fastest model ever. For queries that repeatedly pass the same context, we are increasing the prompt caching discount to 75% (up from 50% previously) for these new models. Finally, we offer long context requests at no additional cost beyond the standard per-token costs.

Model *(Prices are per 1M tokens)*	**Input**	**Cached input**	**Output**	**Blended Pricing***
gpt-4.1	$2.00	$0.50	$8.00	$1.84
gpt-4.1-mini	$0.40	$0.10	$1.60	$0.42
gpt-4.1-nano	$0.10	$0.025	$0.40	$0.12

**Based on typical input/output and cache ratios.*

These models are available for use in our Batch API at an additional 50% pricing discount.

Conclusion

GPT-4.1 is a significant step forward in the practical application of AI. By focusing closely on real-world developer needs—ranging from coding to instruction-following and long context understanding—these models unlock new possibilities for building intelligent systems and sophisticated agentic applications. We're continually inspired

定价

GPT-4.1、GPT-4.1 mini 和 GPT-4.1 nano 现已向所有开发者开放。

通过对我们的推理系统进行效率提升，我们能够为GPT-4.1系列提供更低的价格。GPT-4.1在中位查询中比GPT-4o便宜26%，而GPT-4.1 nano是我们有史以来最便宜、最快的模型。对于反复传递相同上下文的查询，我们将这些新模型的提示缓存折扣提高到75%（之前为50%）。最后，长上下文请求在标准每个令牌费用之外不收取额外费用。

Model *(Prices are per 1M tokens)*	**Input**	**Cached input**	**Output**	**Blended Pricing***
gpt-4.1	$2.00	$0.50	$8.00	$1.84
gpt-4.1-mini	$0.40	$0.10	$1.60	$0.42
gpt-4.1-nano	$0.10	$0.025	$0.40	$0.12

**Based on typical input/output and cache ratios.*

这些模型可以在我们的 中使用，享受额外50%的价格批处理API折扣。

结论

GPT-4.1 在人工智能的实际应用方面迈出了重要的一步。通过紧密关注现实世界中的开发者需求——从编码到指令执行以及长上下文理解——这些模型为构建智能系统和复杂的代理应用开辟了新的可能性。我们不断受到启发

Appendix

A full list of results across academic, coding, instruction following, long context, vision, and function calling evals can be found below.

Academic knowledge

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
AIME '24	48.1%	49.6%	29.4%	13.1%	8.6%	74.3%	87.3%	36.7%
GPQA Diamond[1]	66.3%	65.0%	50.3%	46.0%	40.2%	75.7%	77.2%	69.5%
MMLU	90.2%	87.5%	80.1%	85.7%	82.0%	91.8%	86.9%	90.8%
Multilingual MMLU	87.3%	78.5%	66.9%	81.4%	70.5%	87.7%	80.7%	85.1%

[1] Our implementation of GPQA uses a model to extract the answer instead of regex. For GPT-4.1, the difference was <1% (not statistically significant), but for GPT-4o model extraction improves scores significantly (~46% → 54%).

Coding evals

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
SWE-bench Verified[2]	54.6%	23.6%	-	33.2%	8.7%	41.0%	49.3%	38.0%
SWE-Lancer	$176K (35.1%)	$165K (33.0%)	$77K (15.3%)	$163K (32.6%)	$116K (23.1%)	$160K (32.1%)	$90K (18.0%)	$186K (37.3%)

由开发者提供eloper 社区的创造力，并且很期待看到你用 GPT-4.1 构建的作品。

附录

以下是学术、编码、指令遵循、长上下文、视觉和函数调用评估的完整结果列表。

学术知识

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
AIME '24	48.1%	49.6%	29.4%	13.1%	8.6%	74.3%	87.3%	36.7%
GPQA Diamond[1]	66.3%	65.0%	50.3%	46.0%	40.2%	75.7%	77.2%	69.5%
MMLU	90.2%	87.5%	80.1%	85.7%	82.0%	91.8%	86.9%	90.8%
Multilingual MMLU	87.3%	78.5%	66.9%	81.4%	70.5%	87.7%	80.7%	85.1%

[1] 我们对 GPQA 的实现使用模型来提取答案，而不是正则表达式。对于 GPT-4.1，差异为 <1%（没有统计学意义），但对于 GPT-4o 模型，提取显著提高得分（~46% → 54%）。

编码评估

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
SWE-bench Verified[2]	54.6%	23.6%	-	33.2%	8.7%	41.0%	49.3%	38.0%
SWE-Lancer	$176K (35.1%)	$165K (33.0%)	$77K (15.3%)	$163K (32.6%)	$116K (23.1%)	$160K (32.1%)	$90K (18.0%)	$186K (37.3%)

Category	4.1	4.1 mini	nano	(2024-11-20)	4o mini	(high)	(high)	4.5
SWE-Lancer (IC-Diamond subset)	$34K (14.4%)	$31K (13.1%)	$9K (3.7%)	$29K (12.4%)	$11K (4.8%)	$29K (9.7%)	$17K (7.4%)	$41K (17.4%)
Aider's polyglot: whole	51.6%	34.7%	9.8%	30.7%	3.6%	64.6%	66.7%	-
Aider's polyglot: diff	52.9%	31.6%	6.2%	18.2%	2.7%	61.7%	60.4%	44.9%

[2] We omit 23/500 problems that could not run on our infrastructure. The full list of 23 tasks omitted are 'astropy__astropy-7606', 'astropy__astropy-8707', 'astropy__astropy-8872', 'django__django-10097', 'django__django-7530', 'matplotlib__matplotlib-20488', 'matplotlib__matplotlib-20676', 'matplotlib__matplotlib-20826', 'matplotlib__matplotlib-23299', 'matplotlib__matplotlib-24970', 'matplotlib__matplotlib-25479', 'matplotlib__matplotlib-26342', 'psf__requests-6028', 'pylint-dev__pylint-6528', 'pylint-dev__pylint-7080', 'pylint-dev__pylint-7277', 'pytest-dev__pytest-5262', 'pytest-dev__pytest-7521', 'scikit-learn__scikit-learn-12973', 'sphinx-doc__sphinx-10466', 'sphinx-doc__sphinx-7462', 'sphinx-doc__sphinx-8265', and 'sphinx-doc__sphinx-9367'.

Instruction Following Eval

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
Internal API instruction following (hard)	49.1%	45.1%	31.6%	29.2%	27.2%	51.3%	50.0%	54.0%
MultiChallenge	38.3%	35.8%	15.0%	27.8%	20.3%	44.9%	39.9%	43.8%
MultiChallenge (o3-mini grader)[3]	46.2%	42.2%	31.1%	39.9%	25.6%	52.9%	50.2%	50.1%
COLLIE	65.8%	54.6%	42.5%	50.2%	52.7%	95.3%	98.7%	72.3%
IFEval	87.4%	84.1%	74.5%	81.0%	78.4%	92.2%	93.9%	88.2%
Multi-IF	70.8%	67.0%	57.2%	60.9%	57.9%	77.9%	79.5%	70.8%

[3] Note: we find that the default grader in MultiChallenge (GPT-4o) frequently mis-scores model responses. We find that swapping the grader to a reasoning model, like o3-mini, improves accuracy on grading significantly on samples we've inspected. For consistency reasons with the

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
SWE-Lancer (IC-Diamond subset)	$34K (14.4%)	$31K (13.1%)	$9K (3.7%)	$29K (12.4%)	$11K (4.8%)	$29K (9.7%)	$17K (7.4%)	$41K (17.4%)
Aider's polyglot: whole	51.6%	34.7%	9.8%	30.7%	3.6%	64.6%	66.7%	-
Aider's polyglot: diff	52.9%	31.6%	6.2%	18.2%	2.7%	61.7%	60.4%	44.9%

[2] 我们省略了 23/500 个无法在我们的基础设施上运行的问题。被省略的 23 个任务的完整列表为 'astropy__astropy-7606'、'astropy__astropy-8707'、'astropy__astropy-8872'、'django__django-10097'、'django__django-7530'、'matplotlib__matplotlib-20488'、'matplotlib__matplotlib-20676'、'matplotlib__matplotlib-20826'、'matplotlib__matplotlib-23299'、'matplotlib__matplotlib-24970'、'matplotlib__matplotlib-25479'、'matplotlib__matplotlib-26342'、'psf__requests-6028'、'pylint-dev__pylint-6528'、'pylint-dev__pylint-7080'、'pylint-dev__pylint-7277'、'pytest-dev__pytest-5262'、'pytest-dev__pytest-7521'、'scikit-learn__scikit-learn-12973'、'sphinx-doc__sphinx-10466'、'sphinx-doc__sphinx-7462'、'sphinx-doc__sphinx-8265' 和 'sphinx-doc__sphinx-9367'。

指令执行评估

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
Internal API instruction following (hard)	49.1%	45.1%	31.6%	29.2%	27.2%	51.3%	50.0%	54.0%
MultiChallenge	38.3%	35.8%	15.0%	27.8%	20.3%	44.9%	39.9%	43.8%
MultiChallenge (o3-mini grader)[3]	46.2%	42.2%	31.1%	39.9%	25.6%	52.9%	50.2%	50.1%
COLLIE	65.8%	54.6%	42.5%	50.2%	52.7%	95.3%	98.7%	72.3%
IFEval	87.4%	84.1%	74.5%	81.0%	78.4%	92.2%	93.9%	88.2%
Multi-IF	70.8%	67.0%	57.2%	60.9%	57.9%	77.9%	79.5%	70.8%

[3] 注意：我们发现 MultiChallenge（GPT-4o）中的默认评分器经常误判模型的回答。我们发现将评分器换成一个推理模型，比如 o3-mini，在我们检查的样本中显著提高了评分的准确性。为了保持一致性，

Long Context Evals

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
OpenAI-MRCR: 2 needle128k	57.2%	47.2%	36.6%	31.9%	24.5%	22.1%	18.7%	38.5%
OpenAI-MRCR: 2 needle 1M	46.3%	33.3%	12.0%	-	-	-	-	-
Graphwalks bfs <128k	61.7%	61.7%	25.0%	41.7%	29.0%	62.0%	51.0%	72.3%
Graphwalks bfs >128k	19.0%	15.0%	2.9%	-	-	-	-	-
Graphwalks parents <128k	58.0%	60.5%	9.4%	35.4%	12.6%	50.9%	58.3%	72.6%
Graphwalks parents >128k	25.0%	11.0%	5.6%	-	-	-	-	-

Vision Eval

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
MMMU	74.8%	72.7%	55.4%	68.7%	56.3%	77.6%	-	75.2%
MathVista	72.2%	73.1%	56.2%	61.4%	56.5%	71.8%	-	72.3%
CharXiv-R	56.7%	56.8%	40.5%	52.7%	36.8%	55.1%	-	55.4%
CharXiv-D	87.9%	88.4%	73.9%	85.3%	76.6%	88.9%	-	90.0%

排行榜，我们正在公布这两组结果。注意：我们发现 MultiChallenge（GPT-4o）中的默认评分器经常误判模型{v*}。responses. 我们的回答。将评分器切换到像 o3-mini 这样的推理模型，显著提高我们在样本上的评分准确性检查完毕。对于 出于与排行榜的一致性考虑，我们同时公布了两个结果集。

长上下文评估

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
OpenAI-MRCR: 2 needle128k	57.2%	47.2%	36.6%	31.9%	24.5%	22.1%	18.7%	38.5%
OpenAI-MRCR: 2 needle 1M	46.3%	33.3%	12.0%	-	-	-	-	-
Graphwalks bfs <128k	61.7%	61.7%	25.0%	41.7%	29.0%	62.0%	51.0%	72.3%
Graphwalks bfs >128k	19.0%	15.0%	2.9%	-	-	-	-	-
Graphwalks parents <128k	58.0%	60.5%	9.4%	35.4%	12.6%	50.9%	58.3%	72.6%
Graphwalks parents >128k	25.0%	11.0%	5.6%	-	-	-	-	-

视觉评估

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
MMMU	74.8%	72.7%	55.4%	68.7%	56.3%	77.6%	-	75.2%
MathVista	72.2%	73.1%	56.2%	61.4%	56.5%	71.8%	-	72.3%
CharXiv-R	56.7%	56.8%	40.5%	52.7%	36.8%	55.1%	-	55.4%
CharXiv-D	87.9%	88.4%	73.9%	85.3%	76.6%	88.9%	-	90.0%

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
ComplexFuncBench	65.5%	49.3%	0.6%	66.5%	38.6%	47.6%	17.6%	63.0%
Taubench airline[4]	49.4%	36.0%	14.0%	42.8%	22.0%	50.0%	32.4%	50.0%
Taubench retail[4, 5]	68.0% (73.6%)	55.8% (65.4%)	22.6% (23.5%)	60.3%	44.0%	70.8%	57.6%	68.4%

[4] tau-bench eval numbers are averaged across 5 runs to reduce variance, and run without any custom tools or prompting.

[5] Numbers in parentheses represent Tau-bench results when using GPT-4.1 as the user model, rather than GPT-4o. We've found that, since GPT-4.1 is better at instruction following, it is better able to perform as the user, and so results in more successful trajectories. We believe this represents the true performance of the evaluated model on the benchmark.

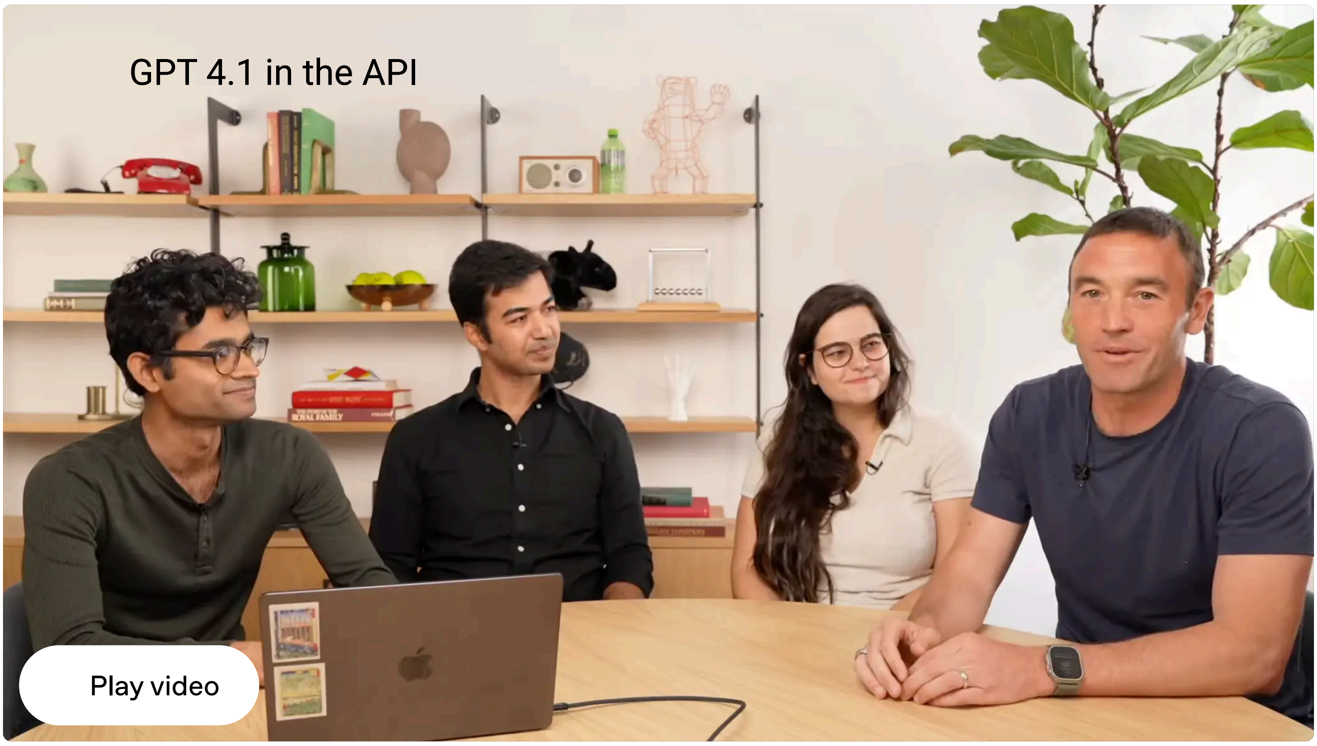
Livestream replay

函数调用 Eval

Category	GPT-4.1	GPT-4.1 mini	GPT-4.1 nano	GPT-4o (2024-11-20)	GPT-4o mini	OpenAI o1 (high)	OpenAI o3-mini (high)	GPT-4.5
ComplexFuncBench	65.5%	49.3%	0.6%	66.5%	38.6%	47.6%	17.6%	63.0%
Taubench airline[4]	49.4%	36.0%	14.0%	42.8%	22.0%	50.0%	32.4%	50.0%
Taubench retail[4, 5]	68.0% (73.6%)	55.8% (65.4%)	22.6% (23.5%)	60.3%	44.0%	70.8%	57.6%	68.4%

[4] tau-bench 评估数字在 5 次运行中取平均以减少方差，且不使用任何自定义工具或提示。 [5] 括号中的数字代表在使用 GPT-4.1 作为用户模型时的 Tau-bench 结果，而非 GPT-4o。我们发现，由于 GPT-4.1 在指令执行方面更优，它更能胜任作为用户，从而产生更多成功的轨迹。我们相信这代表了被评估模型在基准测试中的真实表现。

直播回放



OpenAI

2025 API Platform

Author

OpenAI

Research leads

Ananya Kumar, Jiahui Yu, John Hallman, Michelle Pokrass

Research core contributors

Adam Goucher, Adi Ganesh, Bowen Cheng, Brandon McKinzie, Brian Zhang, Chris Koch, Colin Wei, David Medina, Edmund Wong, Erin Kavanaugh, Florent Bekerman, Haitang Hu, Hongyu Ren, Ishaan Singal, Jamie Kiros, Jason Ai, Ji Lin, Jonathan Chien, Josh McGrath, Julian Lee, Julie Wang, Kevin Lu, Kristian Georgiev, Kyle Luther, Li Jing, Max Schwarzer, Miguel Castro, Nitish Keskar, Rapha Gontijo Lopes, Shengjia Zhao, Sully Chen, Suvansh Sanjeev, Taylor Gordon, Ted Sanders, Wenda Zhou, Yang Song, Yujia Xie, Yujia Jin, Zhishuai Zhang

Research contributors

Aditya Ramesh, Aiden Low, Alex Nichol, Andrei Gheorghe, Andrew Tulloch, Behrooz Ghorbani, Borys Minaiev, Brandon Houghton, Charlotte Cole, Chris Lu, Edmund Wong, Hannah Sheahan, Jacob Huh, James Qin, Jianfeng Wang, Jonathan Ward, Joseph Mo, Joyce Ruffell, Kai Chen, Karan Singhal, Karina Nguyen, Kenji Hata, Kevin Liu, Maja Trębacz, Matt Lim, Mikhail Pavlov, Ming Chen, Morgan Griffiths, Nat McAleese, Nick Stathas, Rajkumar Samuel, Ravi Teja Mullapudi, Rowan Zellers, Shengli Hu, Shuchao Bi, Spencer Papay, Szi-chieh Yu, Yash Patil, Yufeng Zhang

Applied and scaling contributors

Adam Walker, Ali Kamali, Alvin Wan, Andy Wang, Ben Leimberger, Beth Hoover, Brian Yu, Charlie Jatt, Chen Ding, Cheng Chang, Daniel Kappler, Dinghua Li, Felipe Petroski Such, Janardhanan Vembunarayanan, Joseph Florencio, Kevin King, Larry Lv, Lin Yang, Linden Li, Manoli Liodakis, Mark Hudnall, Nikunj Handa, Olivier Godement, Ryszard Madej, Sean Chang, Sean Fitzgerald, Sherwin Wu, Siyuan Fu, Stanley Hsieh, Yunxing Dai

OpenAI

2025 API 平台

作者

OpenAI

研究引领

阿纳尼亚·库马尔，余嘉慧，约翰·霍尔曼，米歇尔·波克拉斯

研究核心贡献者

亚当·高舍尔，阿迪·甘尼什，鲍恩·程，布兰登·麦金齐，布莱恩·张，克里斯·科赫，科林·魏，戴维·梅迪纳，埃德蒙·王，艾琳·卡纳文，弗洛朗·贝克曼，海棠·胡，荣红宇，伊尚·辛格尔，杰米·基罗斯，杰森·艾，纪林，乔纳森·钱，乔什·麦格拉斯，朱利安·李，朱莉·王，凯文·卢，克里斯蒂安·乔治耶夫，凯尔·卢瑟，李靖，马克斯·施瓦泽，米格尔·卡斯特罗，尼蒂什·凯斯卡，拉法·贡蒂霍·洛佩斯，赵胜佳，苏利·陈，苏万什·桑吉夫，泰勒·戈登，泰德·桑德斯，周文达，宋阳，谢雨佳，金雨佳，张志帅

研究贡献者

阿迪蒂亚·拉梅什（Aditya Ramesh）、艾登·洛（Aiden Low）、亚历克斯·尼科尔（Alex Nichol）、安德烈·盖奥尔（Andrei Gheorghe）、安德鲁·图洛奇（Andrew Tulloch）、贝鲁兹·戈尔巴尼（Behrooz Ghorbani）、博瑞斯·米奈耶夫（Borys Minaiev）、布兰登·霍顿（Brandon Houghton）、夏洛特·科尔（Charlotte Cole）、克里斯·卢（Chris Lu）、埃德蒙·王（Edmund Wong）、汉娜·希恩（Hannah Sheahan）、雅各布·胡（Jacob Huh）、詹姆斯·秦（James Qin）、简峰·王（Jianfeng Wang）、乔纳森·沃德（Jonathan Ward）、约瑟夫·莫（Joseph Mo）、乔伊斯·鲁菲尔（Joyce Ruffell）、凯（Kai Chen）、卡兰·辛格尔（Karan Singhal）、卡丽娜·阮（Karina Nguyen）、畔次·畑（Kenji Hata）、凯文·刘（Kevin Liu）、玛雅·特拉巴奇（Maja Tr bacz）、马特·林（Matt Lim）、米哈伊尔·帕夫洛夫（Mikhail Pavlov）、明（Ming Chen）、摩根·格里菲斯（Morgan Griffiths）、纳特·麦克阿利斯（Nat McAleese）、尼克·斯塔萨斯（Nick Stathas）、拉杰库马尔·塞缪尔（Rajkumar Samuel）、拉维·特贾·穆拉普迪（Ravi Teja Mullapudi）、罗恩·泽勒斯（Rowan Zellers）、胡胜利（Shengli Hu）、毕书超（Shuchao Bi）、斯宾塞·帕佩（Spencer Papay）、余子杰（Szi-chieh Yu）、亚什·帕蒂尔（Yash Patil）、张宇峰（Yufeng Zhang）

应用与扩展贡献者

Adam Walker、Ali Kamali、Alvin Wan、Andy Wang、Ben Leimberger、Beth Hoover、Brian Yu、Charlie Jatt、Chen Ding、Cheng Chang、Daniel Kappler、Dinghua Li、Felipe Petroski Such、Janardhanan Vembunarayanan、Joseph Florencio、Kevin King、Larry Lv、Lin Yang、Linden Li、Manoli Liodakis、Mark Hudnall、Nikunj Handa、Olivier Godement、Ryszard Madej、Sean Chang、Sean Fitzgerald、Sherwin Wu、Siyuan Fu、Stanley Hsieh、Yunxing Dai

Andy Wood, Ashley Tyra, Cary Hudson, Dana Palmie, Jessica Shieh, Justin Wang, Karan Sekhri, Katie Kim, Kendal Simon, Laura Peng, Leher Pathak, Lindsay McCallum, Matt Nichols, Nick Pyne, Noah MacCallum, Oona Gleeson, Pranav Deshpande, Rishabh Aggarwal, Scott Ethersmith, Shaokyi Amdo, Stephen Gutierrez, Tabarak Khan, Terry Lee, Thomas Degry, Veit Moeller, Yara Khakbaz

Our Research

Research Index

Research Overview

Research Residency

Latest Advancements

OpenAI o1

OpenAI o1-mini

GPT-4o

GPT-4o mini

Sora

Safety

Safety Approach

Security & Privacy

ChatGPT

Explore ChatGPT

Team

Enterprise

Education

Pricing

Download

Sora

Sora Overview

Features

Pricing

Sora log in ↗

API Platform

Platform Overview

Pricing

For Business

Overview

Company

About us

Our Charter

Careers

Brand

More

News

Stories

Help Center ↗

Terms & Policies

Terms of Use

Privacy Policy

Security

Other Policies

销售额、市场营销、传播与设计 Andy Wood、Ashley Tyra、Cary Hudson、Dana Palmie、Jessica Shen、Justin Wang、Karan Sekhri、

凯蒂·金、肯达尔·西蒙、劳拉·彭、勒赫尔·帕塔克、林赛·麦卡勒姆、马特·尼科尔斯、尼克·派恩、诺亚·麦卡勒姆、乌娜·格里森、普拉纳夫·德什潘德、里沙布·阿加瓦尔、斯科特·伊瑟斯米斯、邵奇·安多、斯蒂芬·古铁雷斯、塔巴拉克·汗、特里·李、托马斯·德格里、费特·穆勒、雅拉·卡赫巴兹

Our Research

Research Index

Research Overview

Research Residency

Latest Advancements

OpenAI o1

OpenAI o1-mini

GPT-4o

GPT-4o mini

Sora

Safety

Safety Approach

Security & Privacy

ChatGPT

Explore ChatGPT

Team

Enterprise

Education

Pricing

Download

Sora

Sora Overview

Features

Pricing

Sora log in ↗

API Platform

Platform Overview

Pricing

For Business

Overview

Company

About us

Our Charter

Careers

Brand

More

News

Stories

Help Center ↗

Terms & Policies

Terms of Use

Privacy Policy

Security

Other Policies

OpenAI

Developer Forum ↗

English United States

OpenAI

API登录文档

开发者论坛 ↗

English United States