



February 12, 2025 Safety Release Milestone

Sharing the latest Model Spec

We've made updates to the Model Spec based on external feedback and our continued research in shaping desired model behavior.

[Read the Model Spec ↗](#)

▶ Listen to article 7:26

🔗 Share

We're sharing a major update to the Model Spec, a document which defines how we want our AI models to behave. This update reinforces our commitments to customizability, transparency, and intellectual freedom to explore, debate, and create with AI without arbitrary restrictions—while ensuring that guardrails remain in place to reduce the risk of real harm. It builds on the foundations we introduced [last May](#), drawing from our experience applying it in varied contexts from alignment research to serving users across the world.

We're also sharing some early results on model adherence with the Model Spec's principles across a broad range of scenarios. These findings highlight progress over time, as well as areas where we can still improve. The Model Spec—like our models—will continue to evolve as we apply it, share it, and listen to feedback from stakeholders. To support broad use and collaboration, we're releasing this version of the Model Spec into the public domain under a Creative Commons CCO license. This means developers and researchers can freely use, adapt, and build on it in their own work.

OpenAI

OpenAI's goal is to create models that are useful, safe, and aligned with the needs of users and developers while advancing our mission to ensure that artificial general intelligence benefits all of humanity. To achieve this goal, we need to iteratively deploy models that empower developers and users, while preventing our models from causing serious harm to our users or others, and maintaining OpenAI's license to operate.

These objectives can sometimes be in conflict, and the Model Spec balances the tradeoffs between them by instructing the model to follow a clearly defined *chain of command*, along with additional principles that set boundaries and default behaviors for various scenarios. This framework prioritizes user and developer control while remaining within clear, well-defined boundaries:

- **Chain of command:** Defines how the model prioritizes instructions from the platform (OpenAI), developer, and user in order. Most of the Model Spec consists of guidelines that we believe are helpful in many cases, but can be overridden by users and developers. This empowers users and developers to fully customize model behavior within boundaries set by platform-level rules.
- **Seek the truth together:** Like a high-integrity human assistant, our models should empower users to make their own best decisions. This involves a careful balance between (1) avoiding steering users with an agenda, defaulting to objectivity while being willing to explore any topic from any perspective, and (2) working to understand the user's goals, clarify assumptions and uncertain details, and give critical feedback when appropriate—requests we've heard and improved on.
- **Do the best work:** Sets basic standards for competence, including factual accuracy, creativity, and programmatic use.
- **Stay in bounds:** Explains how the model balances user autonomy with precautions to avoid facilitating harm or abuse. This new version is intended to be comprehensive, fully covering all the reasons we intend for our models to refuse user or developer requests.
- **Be approachable:** Describes the model's default conversational style—warm, empathetic, and helpful—and how this style can be adapted.
- **Use appropriate style:** Provides default guidance on formatting and delivery. Whether it's neat bullet points, concise code snippets, or a voice conversation,

OpenAI

Upholding intellectual freedom

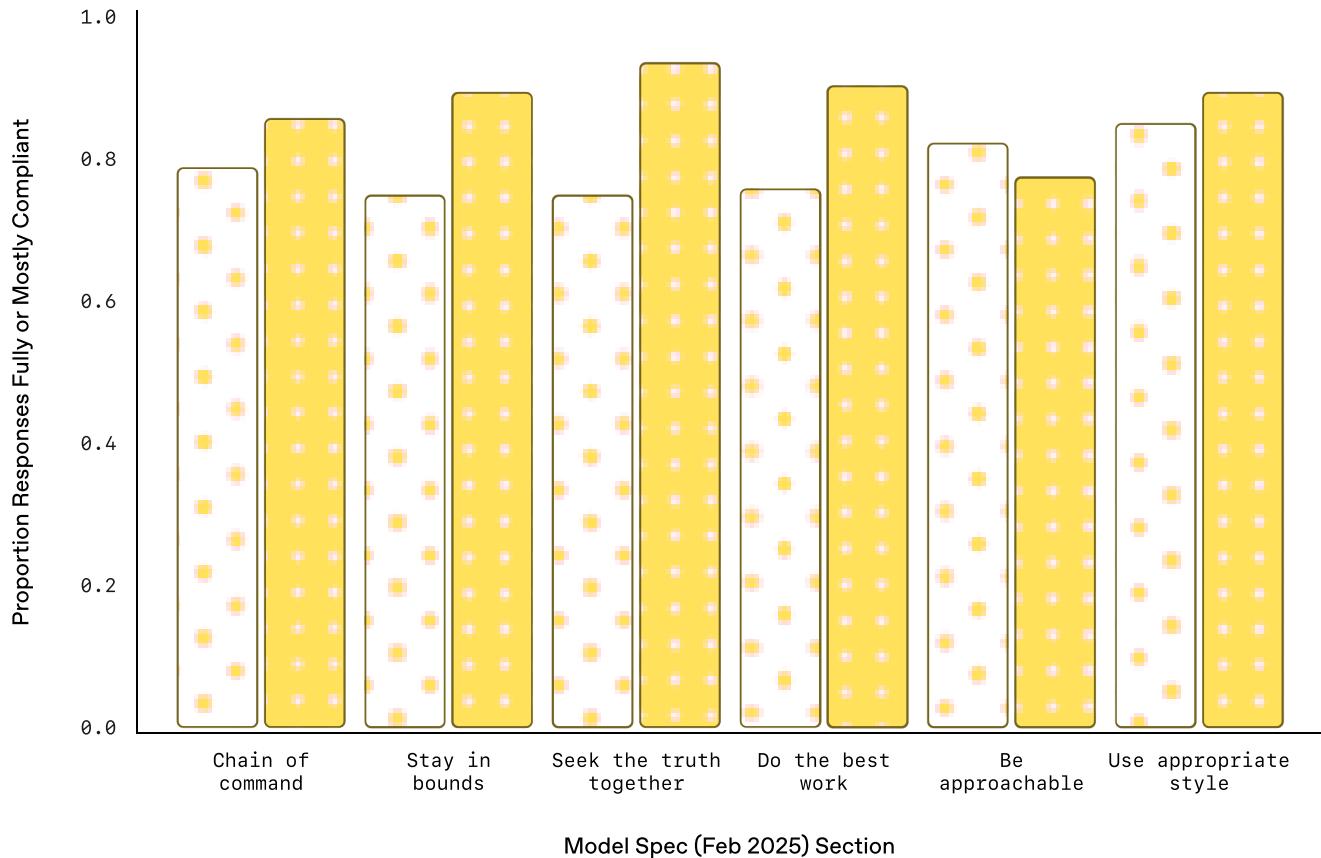
The updated Model Spec explicitly embraces intellectual freedom—the idea that AI should empower people to explore, debate, and create without arbitrary restrictions—no matter how challenging or controversial a topic may be. In a world where AI tools are increasingly shaping discourse, the free exchange of information and perspectives is a necessity for progress and innovation.

This philosophy is embedded in the “Stay in bounds” and “Seek the truth together” sections. For example, while the model should never provide detailed instructions for building a bomb or violating personal privacy, it’s encouraged to provide thoughtful answers to politically or culturally sensitive questions—without promoting any particular agenda. In essence, we’ve reinforced the principle that no idea is inherently off limits for discussion, so long as the model isn’t causing significant harm to the user or others (e.g., carrying out acts of terrorism).

Measuring progress

To better understand real-world performance, we’ve begun gathering a challenging set of prompts designed to test how well models adhere to each principle in the Model Spec. These prompts were created using a combination of model generation and expert human review, ensuring coverage of both typical and more complex scenarios.

OpenAI



Preliminary results show significant improvements in model adherence to the Model Spec compared to our best system last May. While some of this difference may be attributed to policy updates, we believe most of it stems from enhanced alignment. Although the progress is encouraging, we recognize there is still significant room for growth.

We view this as the start of an ongoing process. We plan to keep broadening our challenge set with new examples—especially cases uncovered through real-world use—that our models and the Model Spec do not yet fully address.

In shaping this version of the Model Spec, we incorporated feedback from the first version as well as learnings from alignment research and real-world deployment. In the future we want to consider much more broad public input. To build out processes to that end, we have been conducting pilot studies with around 1,000 individuals—each reviewing model behavior, proposed rules and sharing their thoughts. While these studies are not reflecting broad perspectives yet, early insights directly informed some modifications. We recognize it as an ongoing, iterative process and remain committed to learning and refining our approach.

OpenAI

Open sourcing the Model Spec

We're dedicating this new version of the Model Spec to the public domain under a Creative Commons CC0 license. This means that developers and researchers can freely use, adapt, or build on the Model Spec in their own work. We are also open-sourcing the evaluation prompts used above—and aim to release further code, artifacts, and tools for Spec evaluation and alignment in the future.

You can find these prompts and the Model Spec source in a new [Github repository](#), where we plan to regularly publish new Model Spec versions going forward.

What's next?

As our AI systems advance, we will continue to iterate on these principles, invite community feedback, and openly share our progress. Moving forward, we won't be publishing blog posts for every update to the Model Spec. Instead, you can always find and track the latest updates at model-spec.openai.com.

Our goal is to continuously enable new use cases safely, evolving our approach guided by ongoing research and innovation. AI's growing role in our daily lives makes it essential to keep learning, refining, and engaging openly. This approach reflects not only what we've learned so far but our belief that aligning AI is an ongoing journey—one we hope you'll join us on. If you have feedback on this Spec, you can share it [here](#).

Alignment ChatGPT 2025

Authors

[OpenAI](#)

OpenAI

Our Research	ChatGPT	For Business	Terms & Policies
Research Index	Explore ChatGPT	Overview	Terms of Use
Research Overview	Team	Company	Privacy Policy
Research Residency	Enterprise	About us	Security
	Education	Our Charter	Other Policies
Latest Advancements	Pricing	Careers	
OpenAI o1	Download	Brand	
OpenAI o1-mini			
GPT-4o	Sora	More	
GPT-4o mini	Sora Overview	News	
Sora	Features	Stories	
	Pricing	Help Center ↗	
Safety	Sora log in ↗		
Safety Approach			
Security & Privacy	API Platform		
	Platform Overview		
	Pricing		
	API log in ↗		
	Documentation ↗		
	Developer Forum ↗		

