OpenAI

Menu

Research

Lessons learned on language model safety and misuse

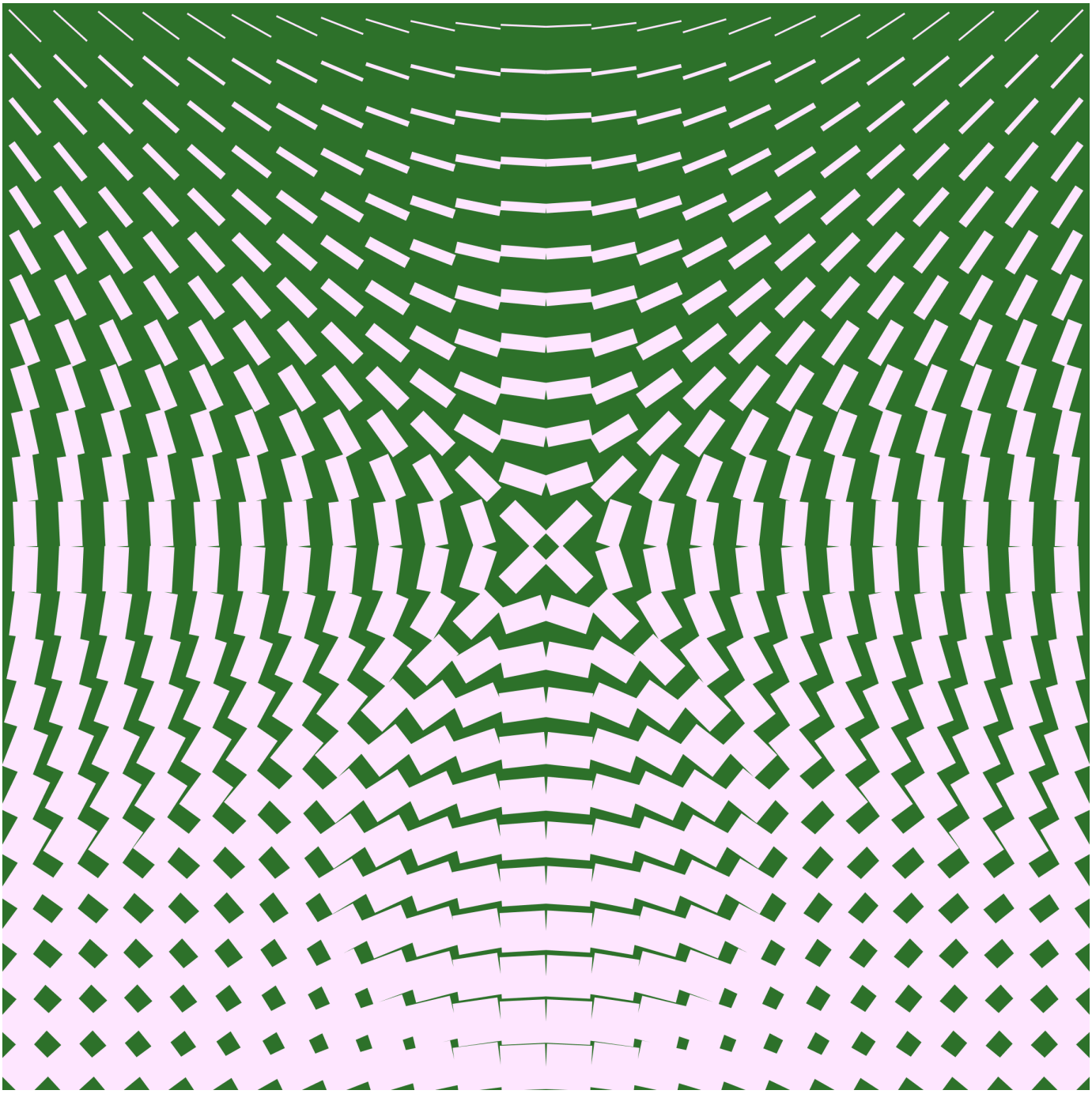


Illustration: Justin Jay Wang

We describe our latest thinking in the hope of helping other AI developers address safety and misuse of deployed models.

March 3, 2022

Read research agenda ↗

Safety & Alignment, Language, Responsible AI, Conclusion

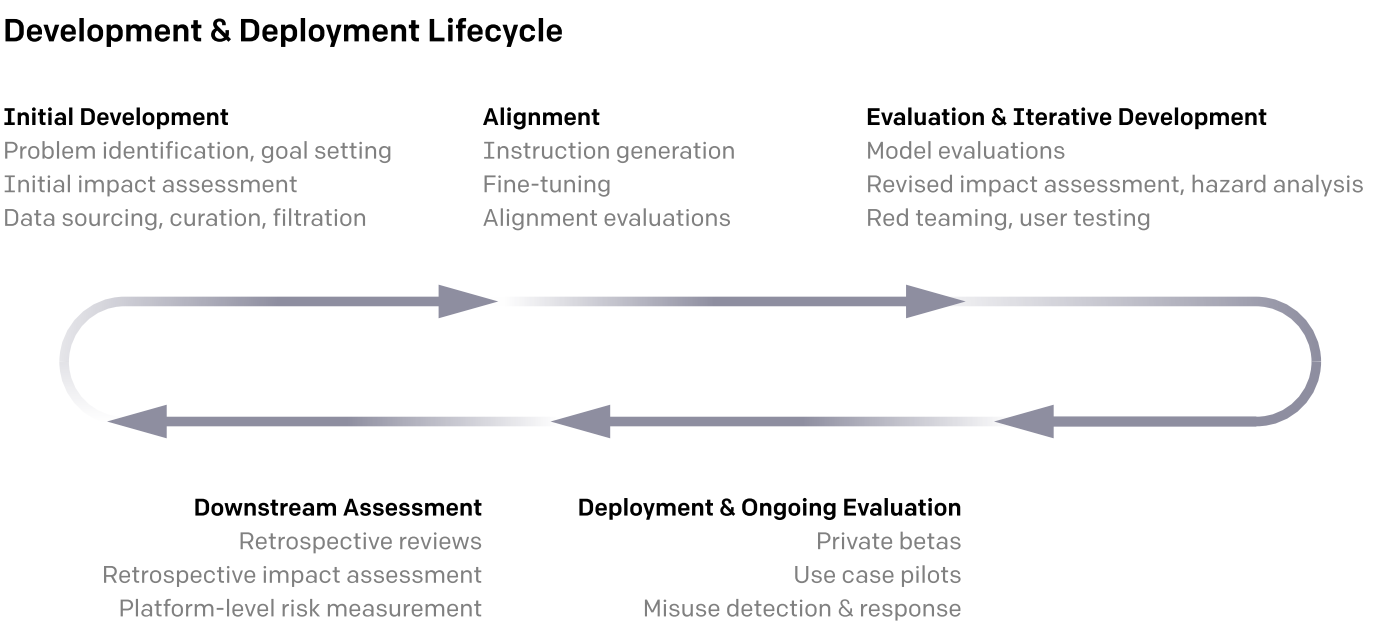
Summary
The deployment of powerful AI systems has enriched our understanding of safety and misuse far more than would have been possible through research alone. Notably: API-based language model misuse often comes in different forms than we feared most; we have identified limitations in existing language model evaluations that we are addressing with novel benchmarks and classifiers; and basic safety research offers significant benefits for the commercial utility of AI systems.

Over the past two years, we've learned a lot about how language models can be used and abused—insights we couldn't have gained without the experience of real-world deployment. In June 2020, we began giving access to developers and researchers to the OpenAI API, an interface for accessing and building applications on top of new AI models developed by OpenAI. Deploying GPT-3, Codex, and other models in a way that reduces risks of harm has posed various technical and policy challenges.

Overview of our model deployment approach

Large language models are now capable of performing a very wide range of tasks, often out of the box. Their risk profiles, potential applications, and wider effects on society remain poorly understood. As a result, our deployment approach emphasizes continuous iteration, and makes use of the following strategies aimed at maximizing the benefits of deployment while reducing associated risks:

- Pre-deployment risk analysis, leveraging a growing set of safety evaluations and red teaming tools (e.g., we checked our InstructGPT for any safety degradations using the evaluations discussed below)
- Starting with a small user base (e.g., both GPT-3 and our InstructGPT series began as private betas)
- Studying the results of pilots of novel use cases (e.g., exploring the conditions under which we could safely enable longform content generation, working with a small number of customers)
- Implementing processes that help keep a pulse on usage (e.g., review of use cases, token quotas, and rate limits)
- Conducting detailed retrospective reviews (e.g., of safety incidents and major deployments)



There is no silver bullet for responsible deployment, so we try to learn about and address our models' limitations, and potential avenues for misuse, at every stage of development and deployment. This approach allows us to learn as much as we can about safety and policy issues at small scale and incorporate those insights prior to launching larger-scale deployments.

There is no silver bullet for responsible deployment.

While not exhaustive, some areas where we've invested so far include[A]:

- Pre-training data curation and filtering
- Fine-tuning models to better follow instructions
- Risk analysis of potential deployments
- Providing detailed user documentation
- Building tools to screen harmful model outputs
- Reviewing use cases against our policies
- Monitoring for signs of misuse
- Studying the impacts of our models

Since each stage of intervention has limitations, a holistic approach is necessary.

There are areas where we could have done more and where we still have room for improvement. For example, when we first worked on GPT-3, we viewed it as an internal research artifact rather than a production system and were not as aggressive in filtering out toxic training data as we might have otherwise been. We have invested more in researching and removing such material for subsequent models. We have taken longer to address some instances of misuse in cases where we did not have clear policies on the subject, and have gotten better at iterating on those policies. And we continue to iterate towards a package of safety requirements that is maximally effective in addressing risks, while also being clearly communicated to developers and minimizing excessive friction.

Still, we believe that our approach has enabled us to measure and reduce various types of harms from language model use compared to a more hands-off approach, while at the same time enabling a wide range of scholarly, artistic, and commercial applications of our models.[B]

The many shapes and sizes of language model misuse

OpenAI has been active in researching the risks of AI misuse since our early work on the malicious use of AI in 2018 and on GPT-2 in 2019, and we have paid particular attention to AI systems empowering influence operations. We have worked with external

experts to develop proofs of concept and promoted careful analysis of such risks by third parties. We remain committed to addressing risks associated with language model-enabled influence operations and recently co-organized a workshop on the subject.[C]

Yet we have detected and stopped hundreds of actors attempting to misuse GPT-3 for a much wider range of purposes than producing disinformation for influence operations, including in ways that we either didn't anticipate or which we anticipated but didn't expect to be so prevalent.[D] Our use case guidelines, content guidelines, and internal detection and response infrastructure were initially oriented towards risks that we anticipated based on internal and external research, such as generation of misleading political content with GPT-3 or generation of malware with Codex. Our detection and response efforts have evolved over time in response to real cases of misuse encountered "in the wild" that didn't feature as prominently as influence operations in our initial risk assessments. Examples include spam promotions for dubious medical products and roleplaying of racist fantasies.

To support the study of language model misuse and mitigation thereof, we are actively exploring opportunities to share statistics on safety incidents this year, in order to concretize discussions about language model misuse.

The difficulty of risk and impact measurement

Many aspects of language models' risks and impacts remain hard to measure and therefore hard to monitor, minimize, and disclose in an accountable way. We have made active use of existing academic benchmarks for language model evaluation and are eager to continue building on external work, but we have also have found that existing benchmark datasets are often not reflective of the safety and misuse risks we see in practice.[E]

Such limitations reflect the fact that academic datasets are seldom created for the explicit purpose of informing production use of language models, and do not benefit from the experience gained from deploying such models at scale. As a result, we've been developing new evaluation datasets and frameworks for measuring the safety of our models, which we plan to release soon. Specifically, we have developed new evaluation metrics for measuring toxicity in model outputs and have also developed in-house classifiers for detecting content that violates our content policy, such as erotic content, hate speech, violence, harassment, and self-harm. Both of these in turn have also been leveraged for improving our pre-training data[F] —specifically, by using the classifiers to filter out content and the evaluation metrics to measure the effects of dataset interventions.

Reliably classifying individual model outputs along various dimensions is difficult, and measuring their social impact at the scale of the OpenAI API is even harder. We have

conducted several internal studies in order to build an institutional muscle for such measurement, but these have often raised more questions than answers.

We are particularly interested in better understanding the economic impact of our models and the distribution of those impacts. We have good reason to believe that the labor market impacts from the deployment of current models may be significant in absolute terms already, and that they will grow as the capabilities and reach of our models grow. We have learned of a variety of local effects to date, including massive productivity improvements on existing tasks performed by individuals like copywriting and summarization (sometimes contributing to job displacement and creation), as well as cases where the API unlocked new applications that were previously infeasible, such as synthesis of large-scale qualitative feedback. But we lack a good understanding of the net effects.

We believe that it is important for those developing and deploying powerful AI technologies to address both the positive and negative effects of their work head-on. We discuss some steps in that direction in the concluding section of this post.

The relationship between the safety and utility of AI systems

In our Charter, published in 2018, we say that we "are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions." We then published a detailed analysis of competitive AI development, and we have closely followed subsequent research. At the same time, deploying AI systems via the OpenAI API has also deepened our understanding of the synergies between safety and utility.

For example, developers overwhelmingly prefer our InstructGPT models—which are fine-tuned to follow user intentions [G] — over the base GPT-3 models. Notably, however, the InstructGPT models were not originally motivated by commercial considerations, but rather were aimed at making progress on long-termalignment problems. In practical terms, this means that customers, perhaps not surprisingly, much prefer models that stay on task and understand the user's intent, and models that are less likely to produce outputs that are harmful or incorrect.[H] Other fundamental research, such as our work onleveraging informationretrieved from the Internet in order to answer questions more truthfully, also has potential to improve the commercial utility of AI systems.[I] These synergies will not always occur. For example, more powerful systems will often take more time to evaluate and align effectively, foreclosing immediate opportunities for profit. And a user's utility and that of society may not be aligned due to negative externalities—consider fully automated copywriting, which can be beneficial for content creators but bad for the information ecosystem as a whole.

It is encouraging to see cases of strong synergy between safety and utility, but we are committed to investing in safety and policy research even when they trade off with commercial utility.

We are committed to investing in safety and policy research even when they trade off against commercial utility.

Ways to get involved

Each of the lessons above raises new questions of its own. What kinds of safety incidents might we still be failing to detect and anticipate? How can we better measure risks and impacts? How can we continue to improve both the safety and utility of our models, and navigate tradeoffs between these two when they do arise?

We are actively discussing many of these issues with other companies deploying language models. But we also know that no organization or set of organizations has all the answers, and we would like to highlight several ways that readers can get more involved in understanding and shaping our deployment of state of the art AI systems.

First, gaining first-hand experience interacting with state of the art AI systems is invaluable for understanding their capabilities and implications. We recently ended the API waitlist after building more confidence in our ability to effectively detect and respond to misuse. Individuals in supported countries and territories can quickly get access to the OpenAI API by signing up here.

Second, researchers working on topics of particular interest to us such as bias and misuse, and who would benefit from financial support, can apply for subsidized API credits using this form. External research is vital for informing both our understanding of these multifaceted systems, as well as wider public understanding.

Finally, today we are publishing a research agenda exploring the labor market impacts associated with our Codex family of models, and a call for external collaborators on carrying out this research. We are excited to work with independent researchers to study the effects of our technologies in order to inform appropriate policy interventions, and to eventually expand our thinking from code generation to other modalities.

If you're interested in working to responsibly deploy cutting-edge AI technologies, apply to work at OpenAI!

Footnotes

A This post is based on our approach to deploying language models through an API, and as such the lessons and mitigations described are most relevant to those also pursuing API-based deployment. However, we also expect some of the discussion to be relevant to those building first-party applications using language models and those considering the open source release of language models. ↩

B This post is intended to explain and share learnings from our approach, rather than to suggest that all actors should necessarily adopt the same approach, or that the same approach is applicable to all possible AI systems. There are benefits and costs associated with different deployment approaches, different models will benefit more or less from study prior to deployment, and in some cases it can be valuable for distinct deployment paths to be pursued by different actors. ↩

C More details on this workshop will be included in the forthcoming publication based on it. ↩

D The mitigations that we emphasize in response to misuse have also evolved. For example, we initially focused on long form text generation as a threat vector, given prior cases of influence operations that involved people manually writing long form misleading content. Given that emphasis, we set maximum output lengths for generated text. Based on a pilot study of long form generation, however, we saw that output restrictions had little effect on policy violations—we've come to believe instead that short-form content amplifying or increasing engagement on misleading content could be the greater risk. ↩

E Examples of limitations in existing datasets, from the perspective of practitioners seeking a holistic assessment of the safety of real language model outputs, include the following: an overly narrow focus (e.g., just measuring occupational gender bias), an overly broad focus (e.g., measuring all under the umbrella of "toxicity"), a tendency to abstract away the specifics of use and context, a failure to measure the *generative* dimension of language model use (e.g., using multiple choice style), prompts that differ stylistically from those typically used in real language model use cases, not capturing dimensions of safety that are important in practice (e.g., an output following or ignoring a safety-motivated constraint in the instruction), or not capturing types of outputs we have found to be correlated with misuse (e.g., erotic content). ↩

F While our efforts are specifically oriented towards addressing limitations in existing benchmarks and in our own models, we also acknowledge that there are limitations to the methods we use such as classifier-based data filtration. For instance, operationally defining the content areas we aim to detect via filtration is challenging and filtration itself can introduce harmful biases. Additionally, the labeling of toxic data is a critical component of this work and ensuring the mental health of these labelers is an industry-wide challenge. ↩

G The relevant "user" of our API may be a developer building an application or an end-user interacting with such an application, depending on context. There are deep questions about the values our aligned models reflect and we hope to build a more nuanced understanding of how to balance the values of wide range of possible users and competing objectives when aligning language models to be more helpful, more truthful and less harmful. ↩

H More aligned models also have more practical advantages such as reducing the need for "prompt engineering" (providing examples of the desired behavior to steer the model in the right direction), saving space in the model's context window which can be used for other purposes.

↩

I Beyond research, we have found that other safety-motivated interventions sometimes have unexpected benefits to customers. For example, rate limits intended to curb spam or misleading content also help customers to control expenses.
↩

Authors

Miles Brundage

Katie Mayer

Tyna Eloundou

Sandhini Agarwal

Steven Adler

Gretchen Krueger

Jan Leike

Pamela Mishkin

Acknowledgments

Thanks to Lilian Weng, Rosie Campbell, Anna Makanju, Bob McGrew, Hannah Wong, Ryan Lowe, Steve Dowling, Mira Murati, Sam Altman, Greg Brockman, Ilya Sutskever, Percy Liang, Peter Welinder, Ethan Perez, Ellie Evans, Helen Ngo, Helen Toner, Justin Jay Wang, Jack Clark, Rishi Bommasani, Girish Sastry, Sarah Shoker, Matt Knight, Bianca Martin, Bob Rotsted, Lama Ahmad, Toki Sherbakov, and others for providing feedback on this post and related work.

Related research

View all research

Building an early warning system for LLM-aided biological threat creation

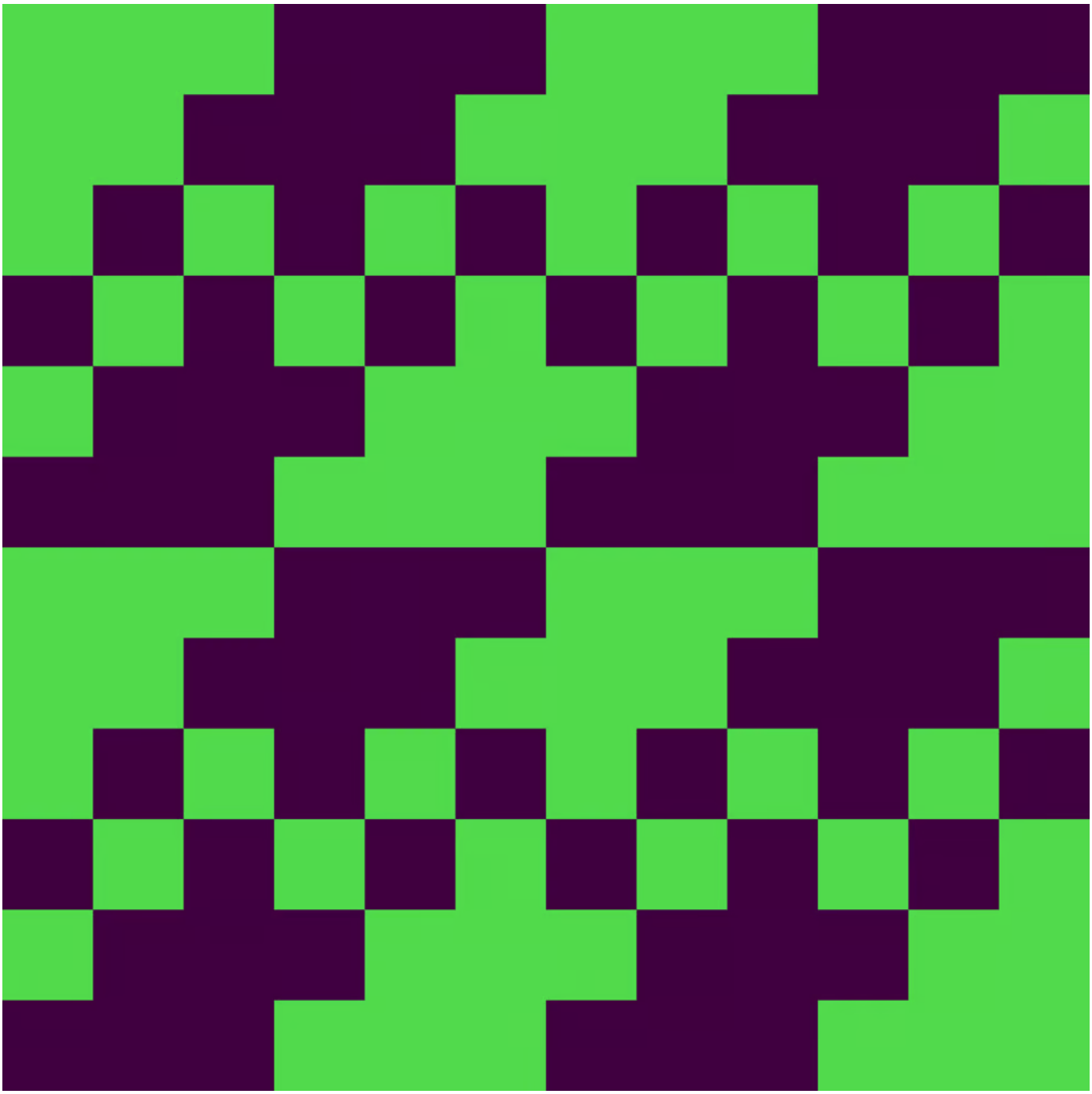
Jan 31, 2024

Weak-to-strong generalization

Dec 14, 2023

Practices for Governing Agentic AI Systems

Dec 14, 2023



GPT-4V(ision) system card

Sep 25, 2023

OpenAI

Research
Overview
Index

API
Overview
Pricing