OpenAI

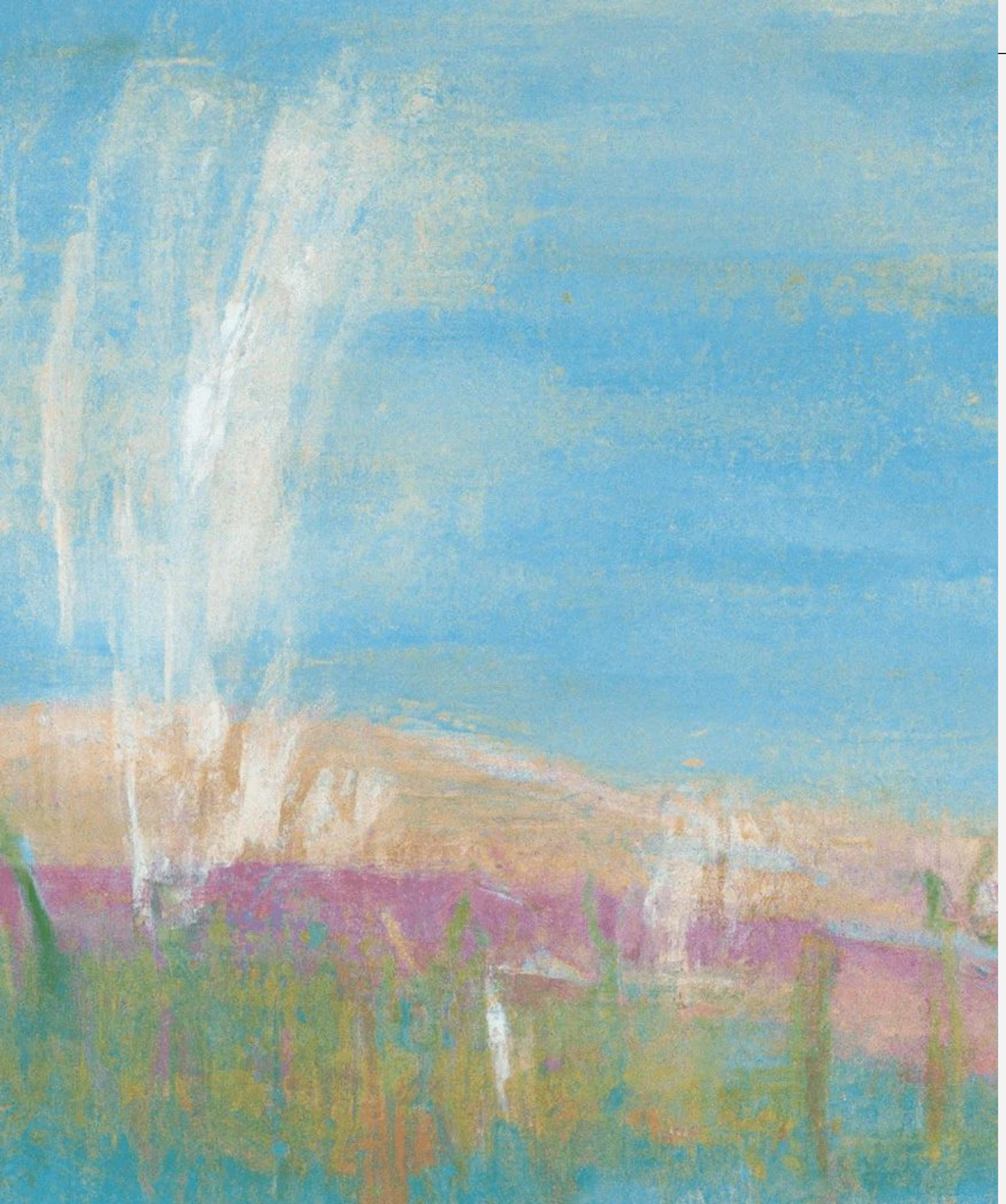
Confidential

How OpenAI is approaching 2024 worldwide elections

May 2024

Introduction

- As we prepare for elections in 2024 across the world’s largest democracies, our approach is to continue our platform safety work by enforcing measured policies, elevating accurate voting information and improving transparency.
- We have a cross-functional effort dedicated to election work, bringing together expertise from our safety systems, threat intelligence, legal, engineering, and policy teams to quickly investigate and address potential abuse.



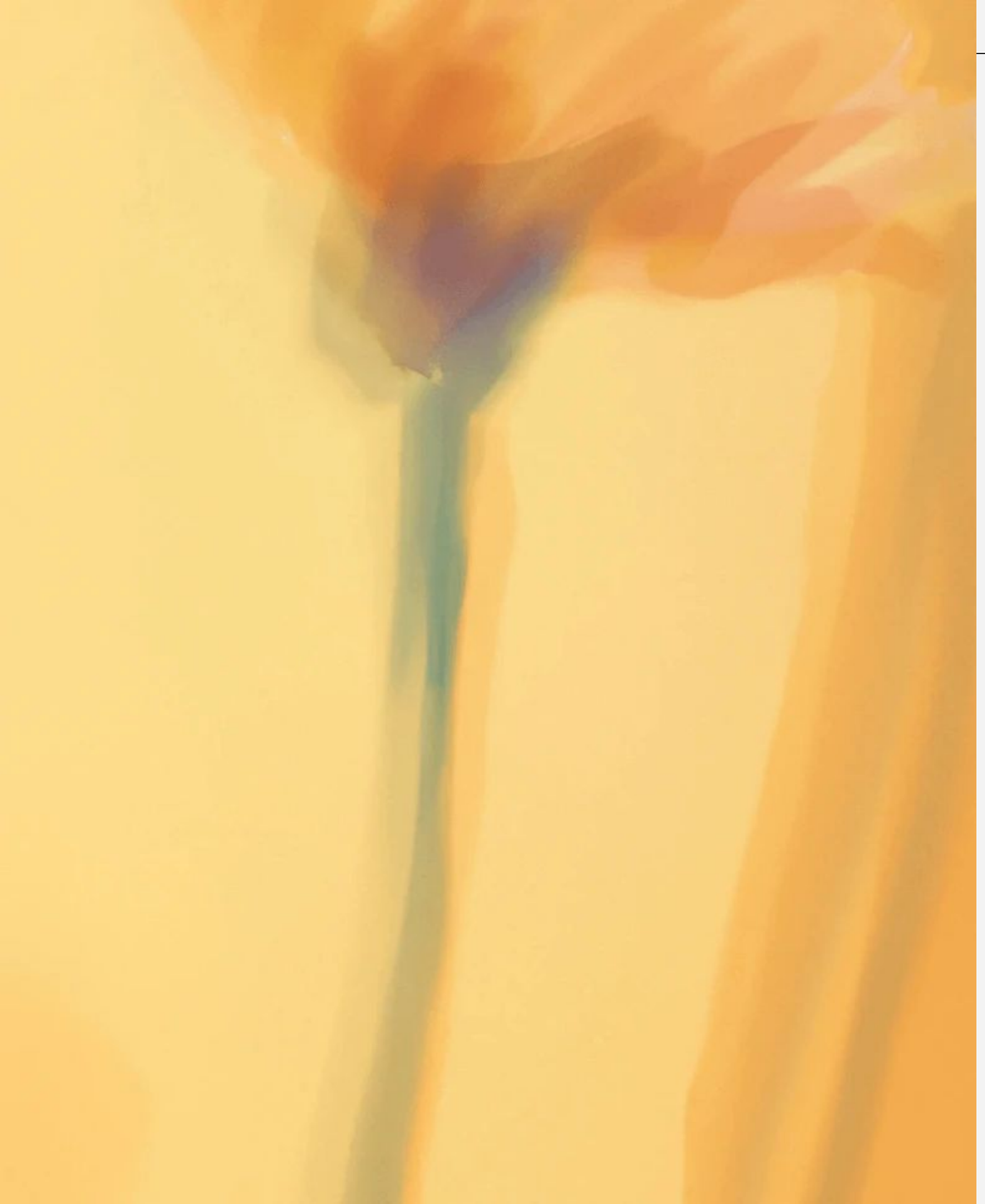
Safety by design

OpenAI’s mission is to build AGI that is **safe** and beneficial for all of humanity.

We design our models with principles such as objectivity in mind - and make these principles public in our **Model Spec.**

Prior to launching new models, we conduct **red teaming** to identify and correct areas of vulnerability, including for topics such as cybersecurity.

Sora has not been publicly released and users can only choose from a preset number of voices on ChatGPT.

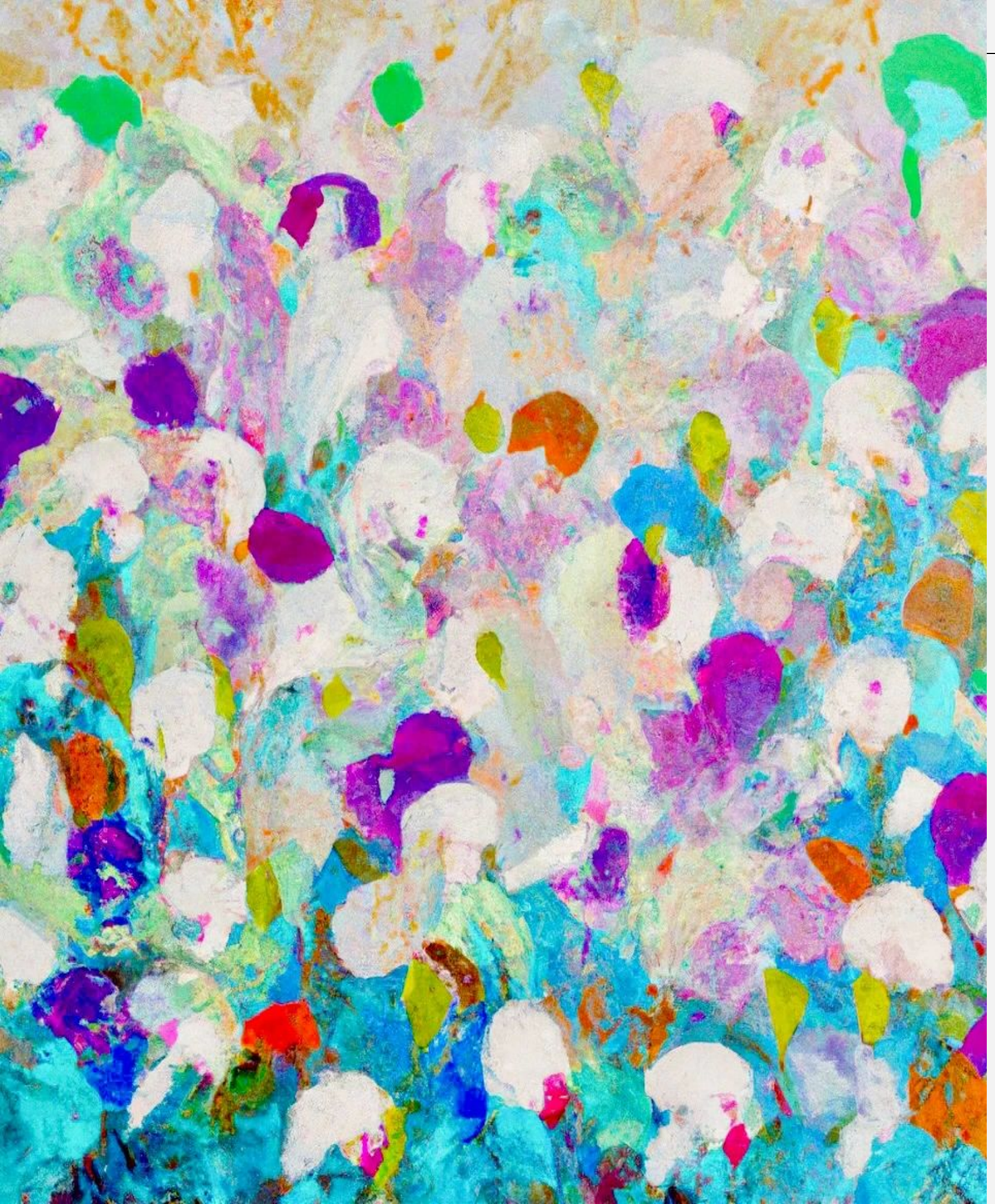


Preventing abuse

Our **usage policy** actively prevents the following from being built on top of our technology: political campaigning and lobbying, impersonation of real people or institutions, democratic interference.

Our **safety monitoring systems** proactively detect unwanted content and with our new GPTs, users can report potential violations to us.

We regularly take down accounts associated with foreign **influence operations.**



Transparency around AI generated content

DALL·E has **guardrails** to decline requests that ask for image generation of real people, including candidates.

We joined the Steering Committee of **C2PA** – the Coalition for Content Provenance and Authenticity and implemented their standards DALL·E 3. We also recently began providing researchers with early access to a new tool that can help **identify images** created by OpenAI's DALL·E 3.

We added **watermarking** into Voice Engine - which is currently open to researchers only.

Elevating accurate information

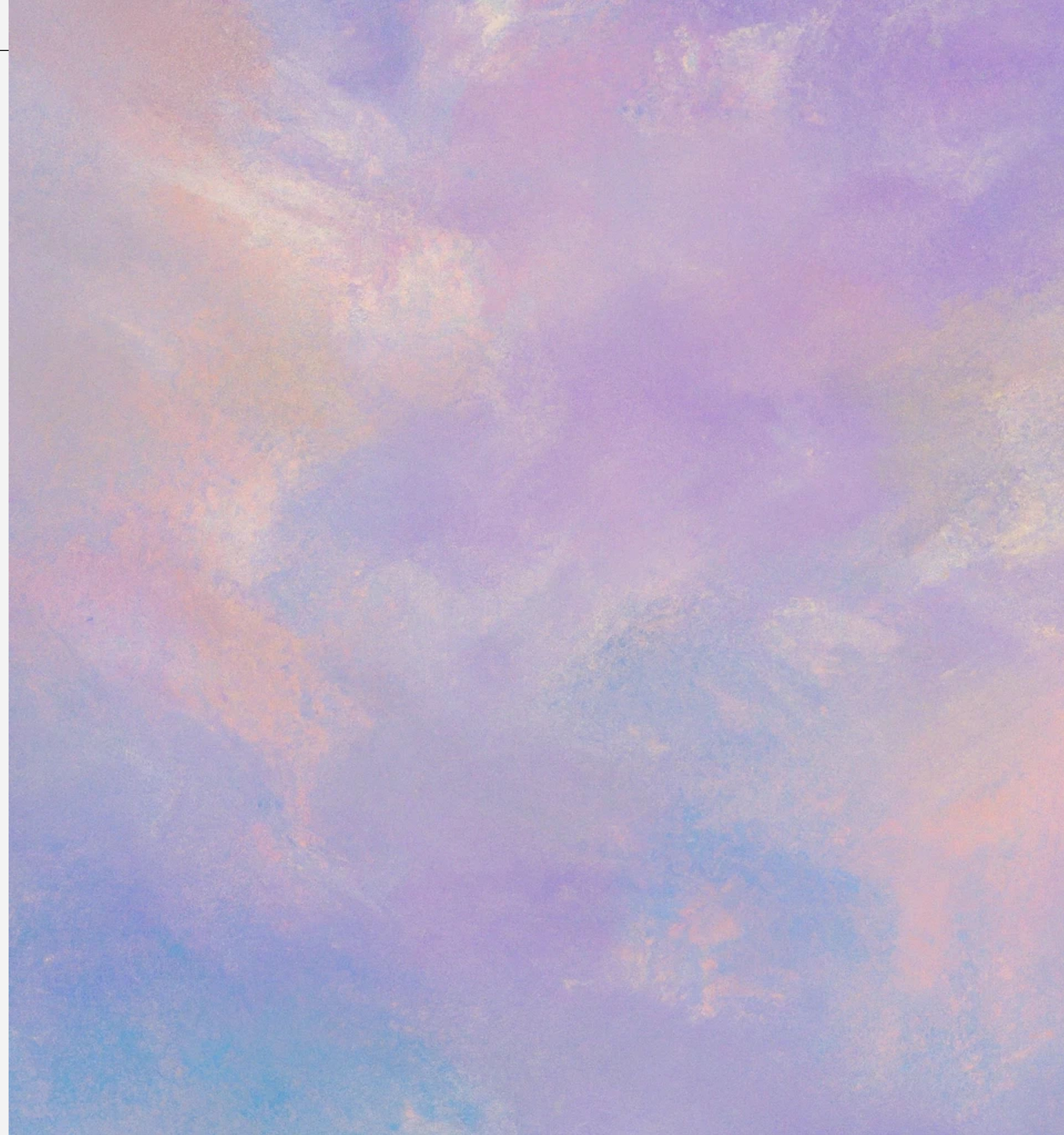
ChatGPT is increasingly integrating with existing sources of information—for example, users will start to get access to real-time news reporting from **AP News** and **Reuters** globally, including attribution and links.

In the US, ChatGPT directs users to CanIVote.org, the authoritative website on US **voting information**, when asked certain procedural election related questions—for example, where to vote. In the EU, ChatGPT redirects to the European Parliament's official website.

Safe by collaboration

We frequently **publish our research into AI safety**, outlining both our approach and key findings from our team. To that end we have launched a grant program to provide $10M for technical safety research tackling issues from weak-to-strong generalization, interpretability, scalable oversight and more.

We announced the creation of the **Societal Resilience Fund** in partnership with Microsoft. With $2M in funding, the Fund will provide AI literacy trainings to election authorities and civil society in key regions around the world.



OpenAI

Confidential

Thank you.