Providing Deep Model Access to Safety Researchers

Anthropic	DeepSeek	Google DeepMind	Meta	OpenAI	x.AI	Zhipu AI
AI Safety researcher Ryan Greenblatt from Redwood Research was recently given employee-level access. [LessWrong, 2023]	Frontier model weights are publicly available	Non-frontier model Gemma 3 model weights publicly available [Google, 2025]	Frontier model weights are publicly available	OpenAI offers a public RL fine-tuning API. [OpenAI]	Non-frontier model Grok-1 model weights are publicly available [xAI, 2024]	Frontier model weights are publicly available

Mentoring and Funding

Anthropic	**Mentoring:** They have their own Anthropic Fellows program and provide a high number of mentors for the independent research seminar program MATS. [Anthropic, 2024; MATS, 2025] **External Researcher Access Program (ongoing):** ▪ gives free API credit to safety/alignment researchers ▪ Standard usage policies apply ▪ $1000 in API Credits (sometimes more) [Anthropic, 2025] **Initiative for developing third-party model evaluations (Jul 2024):** One-off program to provide funding for a third-party to develop evaluations that can effectively measure advanced capabilities in AI models: "The approach is designed to enable you to distribute your evaluations to governments, researchers, and labs focused on AI safety." [Anthropic, 2024].
DeepSeek	None found
Google DeepMind	**Mentoring:** Provides a high number of mentors for the independent research seminar program MATS. [MATS, 2025; MATS]
Meta	None found
OpenAI	**Mentoring:** Currently provides one mentor for the independent research seminar program, MATS. **Researcher Access Program (back since February 2025):** ▪ gives free API credit to safety/alignment researchers ▪ Standard usage policies apply ▪ Up to $1,000 of API credits [OpenAI, 2025] **Superalignment Fast Grants (2023):** $10M to support technical research towards the alignment and safety of superhuman AI systems, including weak-to-strong generalization, interpretability, scalable oversight, and more [OpenAI, 2023].
x.AI	None found
Zhipu AI	None found