| Anthropic | OpenAI | Google DeepMind | Meta | xAI | DeepSeek | Z.ai | Alibaba Cloud |
|---|---|---|---|---|---|---|---|
| AI Safety researcher Ryan Greenblatt from Redwood Research was given employee-level access in 2024, leading to the published research titled "Alignment Faking in Large Language Models." | Non-frontier model gpt-oss-120b and gpt-oss-20b model weights publicly available | Non-frontier model Gemma 3 model weights publicly available | Frontier model weights are publicly available | Non-frontier model Grok-1 model weights are publicly available | Frontier model weights are publicly available | Frontier model weights are publicly available | Non-frontier model Qwen3 model weights are publicly available |