
RED30 AI Red Lines Tracker: A Comprehensive Technical Infrastructure for Monitoring Frontier Model Proximity to Critical Safety Thresholds

Kunal Singh
Independent

Rujuta Karekar
Independent

Aman Agarwal
Independent

With
Apart Research

Abstract

*Each frontier developer publishes self-assessments under its own risk framework: OpenAI's Preparedness Framework, Anthropic's Responsible Scaling Policy, Google DeepMind's Frontier Safety Framework, and etc. However, these assessments remain disconnected, lack a basis tracking for dangerous capabilities, and are rarely presented in a way that allows direct cross-lab comparison or contextualization with compute infrastructure. We present a **web-based dashboard "AI Red Lines Tracker"** to visually track global AI infrastructure and risk landscape in terms of frontier lab's risk analysis, AI RnD acceleration tracker, current proximity to thresholds, areas of convergence and divergence and also tracking training compute, datacentres, compute concentration across frontier labs. We also propose **RED30**, a global baseline framework that defines 30 minimum, non-negotiable safety and ethical boundaries that frontier AI models must not cross. These indicators are derived directly from binding international law, human rights conventions, data protection regulations, criminal law statutes and consumer protection frameworks. We systematically analyse all frontier models with this **RED30** framework and present the findings.*

1. Introduction

AI Models now demonstrate increasingly sophisticated performance across domains previously considered human-exclusive, including scientific reasoning, code generation, strategic planning, and multimodal understanding. While these advances offer immense societal benefit, they simultaneously lower barriers to misuse in ways that could result in catastrophic outcomes [14,15].

Four domains are of particular concern and are explicitly tracked by leading laboratories:

- *Biological and chemical weapons development assistance*
- *Offensive cybersecurity at scale*
- *Large-scale persuasion, manipulation, and disinformation*
- *Autonomous AI research and recursive self-improvement*

Each major developer has responded by publishing structured risk assessments: OpenAI via its Preparedness Framework [1,2], Anthropic through its Responsible Scaling Policy [8], Google DeepMind via the Frontier Safety Framework [10], and xAI via its Frontier AI Framework introduced in late 2025. Despite these efforts, several critical barriers to effective oversight remain.

Our main contributions are:

1. *A visually interactive dashboard showing, AI red lines analysis, frontier lab models risk analysis, compute resources and infrastructure analysis and AI incidents dashboard*
2. ***RED 30***, a standardized framework of ***30 universal AI red line indicators***, organized into 4 categories: Critical Harm (8), Systemic Harm (8), Individual Harm (8), Emerging Standards (6). These indicators serve as priority indicators to measure models risks across categories derived from international regulatory frameworks around the world
3. *AI R&D Tracker showing cross labs comparison to critical thresholds with METR benchmark integrations*
4. *16 frontier model risk analysis across 4 major labs and processed 24 model system cards to show model risk thresholds across CBRN domains.*
5. *Compute Infrastructure dashboard showcasing 20 frontier models training compute (2023-2026), mapping 18 major data centres with 2620 GW global capacity and list of EU AI Act compliant models based on FLOP thresholds.*
6. *Aggregated **354 frontier-model-relevant incidents** from AI Incident Database, categorized incidents by organization and harm types: 15 Offenses, 199 Misuses, 43 Biases, 97 Harmful Outputs*

2. Related Work

Risk assessment frameworks published by frontier laboratories form the primary data foundation of this project [1–11]. OpenAI’s Preparedness Framework [2] provides a detailed per-category scoring and threshold-proximity metrics to date. Anthropic’s Responsible Scaling Policy [8] introduces precautionary ASL tiers, while Google DeepMind’s Frontier Safety Framework [10] emphasizes Capability Confidence Levels and graded alert thresholds. xAI’s published Frontier AI Framework employs quantitative benchmarks across multiple dual-use domains. Independent transparency and evaluation efforts include the Stanford AI Index annual reports [11,19], Epoch AI’s compute and capability trend analysis [12], the incident database [13], METR’s red-line benchmark definitions [16], and Apollo Research’s public model evaluations [16]. Academic literature has explored extreme-risk evaluation methodologies [14] and proposed governance approaches for frontier systems [15]. Civil society initiatives such as the Global Call for AI Red Lines [17] have advocated for clearer thresholds and stronger deployment safeguards.

3. Methods

3.1 Data Acquisition and Extraction

All data is sourced exclusively from publicly released documents: system cards, framework descriptions, blog posts, and third-party benchmark reports.

- *Model metadata (name, release date, system card URL)*
- *Risk scores (no. of models, where available)*
- *Risk levels (lab-specific: Low/Medium/High/Critical, ASL-1–4, CCL stages, etc.)*
- *Key findings*
- *Source document section references*

3.2 Data Processing and Visualization

Chart data preparation occurs in chart-transformers.ts. All visualizations use Recharts (line, bar, radar, heatmap) or react-simple-maps (world map). Lab-specific terminology and color scales are preserved where possible, with optional cross-lab normalization views.

4. Results

The web based dashboard currently tracks 17+ frontier models across four laboratories (OpenAI $\approx$ 6, Anthropic 5, Google DeepMind 3, xAI 4).

Key observations (as of January 2026 data):

- *OpenAI o3 maintains medium rating in biological/chemical threats*
- *xAI Grok 4 family demonstrates particularly strong quantitative performance across biology, chemistry, and cyber benchmarks*

- *Anthropic Claude 4 series models approach preemptive ASL-3 deployment, nearing ASL-4 thresholds in autonomy and CBRN domains*
- *Google DeepMind Gemini 3 series meets alert thresholds in cyber-uplift but remains below them in ML R&D automation*

5. Discussion and Limitations

Implications

The results demonstrate that frontier models from multiple independent laboratories are approaching and in several cases reaching the high-risk or critical red lines defined in their own frameworks. This convergence across organizations strengthens the case for external red-teaming, compute governance, standardized reporting formats, and potential deployment restrictions when clear thresholds are crossed [15,17].

Limitations

- *All data originates from voluntary frameworks inherits potential optimism bias, incomplete evaluation suites, or deliberate minimization*
- *Manual extraction process limits update frequency and introduces human error risk*
- *Direct comparison across frameworks remains imperfect due to differing methodologies and category definitions*
- *No automated ingestion of new documents (planned for future work)*
- *Scope currently limited to four leading laboratories (Open Source for future work)*
- *RED30 framework to measure AI Red lines risk can be extended (Future scope)*
- *AI Incidents dashboard's scope could not be expanded because of limited time, we plan to plot AI incident reports across model release timings, countries and other parameters.*

6. Conclusion

The RED30: AI Red Lines Tracker represents a step toward democratized, and visually accessible monitoring of frontier AI risk assessments. By integrating official self-reports with global compute context, incident tracking, and governance signals, it provides a picture of how close current models are to crossing critical safety thresholds across multiple independent developers.

Code and Data

- **Code repository:** <https://github.com/KUNALSINGH9373/AI-Red-Lines-Tracker-2026>
- **WebSite Dashboard:** <https://ai-red-lines-tracker-2026.vercel.app>

References

1. OpenAI. (2023). *Preparedness Framework (v1)*.
<https://cdn.openai.com/openai-preparedness-framework-beta.pdf>
2. OpenAI. (2024). *Preparedness Framework (v2)*. <https://openai.com/preparedness>
3. OpenAI. (2024). *GPT-4o System Card*. <https://openai.com/index/gpt-4o-system-card/>
4. OpenAI. (2024). *o1 System Card*. <https://openai.com/index/openai-o1-system-card/>
5. OpenAI. (2025). *o3 & o4-mini System Card*.
<https://openai.com/index/o3-o4-mini-system-card/>
6. OpenAI. (2025). *o3-mini System Card*. <https://openai.com/index/o3-mini-system-card/>
7. OpenAI. (2025). *GPT-5 System Card*. <https://openai.com/index/gpt-5-system-card/>
8. Anthropic. (2024). *Responsible Scaling Policy v2.2*.
<https://www.anthropic.com/responsible-scaling-policy>
9. Anthropic. (2025). *Claude 4 family safety reports*. <https://www.anthropic.com/news>
10. Google DeepMind. (2025). *Frontier Safety Framework v3.0*.
<https://deepmind.google/discover/blog/frontier-safety-framework/>
11. Stanford HAI. (2025). *AI Index Report 2025*. <https://aiindex.stanford.edu/report/>
12. Epoch AI. (2025). *Trends in Machine Learning Compute*. <https://epochai.org/trends>
13. *AI Incident Database*. <https://incidentdatabase.ai/>
14. Shevlane et al. (2023). *Model Evaluation for Extreme Risks*. [arXiv:2305.15324](https://arxiv.org/abs/2305.15324)
15. Anderljung et al. (2023). *Frontier AI Regulation*. [arXiv:2307.03718](https://arxiv.org/abs/2307.03718)
16. METR. (2025). *Public evaluation reports*. <https://metr.org>
17. *Global Call for AI Red Lines*. <https://red-lines.ai>

Appendix: Limitations & Dual-Use Considerations

A.1 Methodological Limitations

- *False positives/negatives inherent in self-reported assessments*
- *Manual extraction introduces transcription errors and update delays*
- *Framework incompatibility limits perfect cross-lab quantitative comparison*
- *No independent capability verification performed*

A.2 Dual-Use Risks

The tool could help identify currently highest-risk models in specific domains, potentially informing misuse targeting. All displayed information is already public; no new attack vectors are revealed.

A.3 Responsible Disclosure

No novel vulnerabilities discovered. Future findings would be reported privately to the relevant lab first, with escalation to trusted third parties (AISIs, METR, Apollo) if necessary.

A.4 Ethical Considerations

Prioritizes public-interest transparency over potential information-hazard concerns. Open-source license supports broad access and community auditing.

A.5 Suggested Improvements

- *Automated system-card parsing*
- *Real-time threshold-crossing notifications*
- *Historical version snapshots*
- *Support for additional laboratories*

LLM Usage Statement

***Grok** (xAI) and **Claude** (Anthropic) were used to scrape the data from pdf of model cards files and build a structure representation of the unstructured data. **Claude Code** was used to build the web based dashboard to show all the comparison metrics. All facts, data references, findings, and claims were independently verified against the repository contents (README.md, JSON files, code files, and RED30_Research_Document.md). No data is fabricated.*