
Frontier AI Risk Threshold Analyzer

Swaleha Parveen
AI Safety UAE, Dubai

With
Apart Research

Abstract

The proliferation of AI safety frameworks across laboratories has created a critical governance challenge: different organizations define "dangerous AI" using incompatible taxonomies. Anthropic uses ASL levels, Google DeepMind employs CCL tiers, OpenAI maintains preparedness levels, and the EU AI Act introduces systemic risk thresholds yet no standardized mapping exists between these frameworks. This fragmentation undermines international cooperation, regulatory compliance, and risk assessment consistency.

We present the Frontier AI Risk Threshold Analyzer, an open-source tool that addresses this gap through AI-powered extraction and harmonization of risk thresholds across 12+ major safety frameworks. Using multi-model consensus extraction with GPT-4o, we processed policy documents from Anthropic, OpenAI, DeepMind, and others to build the first comprehensive, machine-readable database of capability thresholds, compute limits, and evaluation requirements. The system enables real-time risk assessment of AI models across all frameworks simultaneously, automated EU AI Act compliance checking, and identification of regulatory arbitrage risks.

Our tool successfully mapped 47 distinct risk tiers across frameworks, revealing that models triggering ASL-3 at Anthropic may not trigger equivalent thresholds at other labs—a critical finding for international coordination. The system is deployed as an interactive web application and has immediate applications in

regulatory compliance, international treaty verification, and responsible scaling policy implementation.

Keywords: AI Governance, Safety Frameworks, Risk Assessment, Compliance Infrastructure, International Coordination, Technical Governance

1. Introduction

1.1 The Framework Fragmentation Problem

As frontier AI systems approach and exceed capabilities that could pose societal risks, major AI laboratories have independently developed safety frameworks to guide responsible development. These frameworks establish "risk tiers" that trigger specific safeguards, evaluations, and deployment restrictions when models reach certain capability or compute thresholds.

However, each organization has developed its own taxonomy:

- Anthropic's Responsible Scaling Policy defines ASL-2 and ASL-3 levels based on CBRN and autonomy risks
- Google DeepMind's Frontier Safety Framework uses CCL (Critical Capability Level) tiers
- OpenAI's Preparedness Framework categorizes models as Low, Medium, High, or Critical risk
- The EU AI Act designates models exceeding 10^{25} FLOPS as "systemic risk" GPAI models under Article 51

This fragmentation creates four critical problems:

1. International Coordination Failure: Countries cannot negotiate meaningful agreements when they cannot compare risk levels across labs. What does it mean when one lab's model is "ASL-3" while another's is "CCL-2"? Are these equivalent risks?

2. Regulatory Compliance Ambiguity: The EU AI Act requires specific actions for "systemic risk" models, but labs using different frameworks struggle to determine compliance status objectively.

3. Regulatory Arbitrage Risk: Labs might exploit framework differences to avoid stricter requirements, choosing to develop models that stay below thresholds in permissive frameworks while exceeding them in stricter ones.

4. Research Gap: No comprehensive, machine-readable database of risk thresholds exists, forcing policymakers and researchers to manually parse policy documents.

1.2 Research Questions

This project addresses three primary questions:

RQ1: Can AI-powered extraction create a comprehensive, structured database of risk thresholds from frontier AI safety policies?

RQ2: To what extent do different frameworks agree or disagree on what constitutes "high-risk" AI, and where are the critical gaps?

RQ3: Can we build practical infrastructure that enables real-time risk assessment and compliance checking across multiple frameworks simultaneously?

1.3 Contributions

Our work makes four primary contributions:

1. **First Comprehensive Database:** We created the first open-source, machine-readable database mapping risk thresholds, capability definitions, and evaluation requirements across 12 major AI lab frameworks.

2. **Multi-Model Extraction Methodology:** We developed and validated a multi-pass AI extraction approach using GPT-4o that achieves high accuracy in parsing complex policy documents.

3. **Interactive Assessment Tool:** We built a deployed web application enabling instant risk assessment of AI models across all frameworks, with automated EU AI Act compliance checking.

4. **Gap Analysis:** We systematically identified where frameworks diverge, revealing potential regulatory arbitrage opportunities and areas requiring international harmonization.

This work directly supports the EU AI Act's Code of Practice development, international AI safety institute coordination, and responsible scaling policy implementation across laboratories.

2. Methods

2.1 Overall Architecture

The Frontier AI Risk Threshold Analyzer consists of three main components:

1. **Data Acquisition Layer:** Automated downloading and processing of policy documents

2. **AI Extraction Pipeline:** Multi-model consensus-based extraction of structured data

3. Assessment & Visualization Layer: Interactive web interface for risk analysis

2.2 Data Collection

We identified and collected policy documents from 12 organizations:

- Anthropic (Responsible Scaling Policy v2.2)
- OpenAI (Preparedness Framework v2)
- Google DeepMind (Frontier Safety Framework v3.0)
- Meta, Microsoft, Amazon, NVIDIA, xAI, Cohere, G42, Naver, Magic

Additionally, we processed:

- EU AI Act Code of Practice documentation (Georgetown CSET analysis)
- Compute threshold research (Institute for Law & AI)
- METR framework analysis

Documents were obtained via direct download (PDFs) or web scraping (HTML pages), totaling approximately 500+ pages of policy text.

2.3 Multi-Model Extraction Pipeline

Traditional single-pass LLM extraction suffers from hallucination and inconsistent interpretation. We developed a multi-pass consensus approach:

Phase 1: Parallel Extraction

- Extract from same document using 3 different prompts with GPT-4o
- Each prompt emphasizes different aspects (capability thresholds, compute limits, evaluations)
- Temperature set to 0.1 for consistency

Phase 2: Consensus Building

- Compare extractions across all three passes
- Identify information appearing in 2+ extractions (high-confidence)
- Flag contradictions for manual review

Phase 3: Reconciliation

- Use GPT-4o to synthesize final structured output
- Prompt: "Reconcile these extractions, include only high-confidence information"

Phase 4: Validation

- Manual spot-checking of 20% of extractions against source documents

- Accuracy rate: 94% for threshold values, 97% for organization names

2.4 Data Schema

All extracted data follows a standardized JSON schema. This enables programmatic access and cross-framework comparisons.

2.5 Assessment Engine

The threshold matcher implements three core algorithms:

Algorithm 1: Framework Mapping

- Input: Model specifications (compute, capabilities, evaluations)
- Process: Match against each framework's thresholds
- Output: Risk tier assignment per framework

Algorithm 2: EU Compliance Check

- Check: Training compute $\geq 10^{25}$ FLOPS?
- If yes: Systemic risk status, list Article 51 requirements
- If no: Below threshold, minimal requirements

Algorithm 3: Gap Analysis

- Compare risk tier assignments across frameworks
- Identify discrepancies (e.g., ASL-3 but not CCL-3)
- Flag potential regulatory arbitrage scenarios

2.6 Implementation Stack

Technology Stack:

- Backend: Python 3.10, FastAPI
- AI: OpenAI API (GPT-4o, gpt-4o-mini)
- Frontend: Streamlit
- Data: JSON + pandas
- Visualization: Plotly

- Deployment: Replit (Autoscale deployment)

Total development time: 48 hours (hackathon duration)

Lines of code: ~2,000

2.7 GitHub Repository

All code is open-source and available at:

<https://github.com/swalehparvin/-Frontier-AI-Risk-Threshold-Analyzer/tree/main>

3. Results



Figure 1 – Representation of compute and framework

Figure 2 – Framework Risk Heatmap

4. Discussion and Conclusion

4.1 Key Findings

This project demonstrates three critical results:

Finding 1: Framework Fragmentation is Severe

We documented 47 distinct risk tiers across 12 frameworks with minimal standardization. This confirms the governance gap identified in policy literature but quantifies it for the first time.

Finding 2: Compute Thresholds are Insufficient

Only the EU AI Act relies primarily on compute (10^{25} FLOPS). Other frameworks emphasize capabilities, creating scenarios where:

- High-compute, safe models over-trigger regulations (false positives)
- Low-compute, dangerous models under-trigger safeguards (false negatives)

As algorithmic efficiency improves $\sim 30\%$ annually (per Epoch AI data), compute thresholds will become increasingly unreliable.

Finding 3: AI-Powered Governance Infrastructure Works

Multi-model extraction achieved 94% accuracy, demonstrating that AI can reliably process policy documents at scale. This enables:

- Automatic monitoring of framework changes
- Real-time compliance checking
- Scalable policy analysis

4.2 Implications for AI Governance

Our findings have three direct policy implications:

Implication 1: International Treaties Need Common Language

The G7, AI Safety Summits, and bilateral agreements cannot succeed without harmonized risk definitions. Our database provides the foundation for developing that common language.

Recommendation: International AI Safety Institutes should adopt standardized risk tier definitions mapped to specific capability thresholds (not just compute).

Implication 2: EU AI Act Needs Capability-Based Refinement

The 10^{25} FLOP threshold is objective and verifiable but insufficient. A 10^{26} FLOP model trained only on public domain text poses different risks than a 10^{24} FLOP model fine-tuned on biological warfare data.

Recommendation: EU Code of Practice should incorporate capability-based triggers alongside compute thresholds.

Implication 3: Regulatory Arbitrage is Possible

We identified scenarios where labs could exploit framework differences. Example:

- Develop model just below ASL-3 capability threshold
- Market in jurisdictions using weaker frameworks
- Avoid strongest safeguards while technically "compliant"

Recommendation: International coordination should establish "red lines" that apply universally regardless of framework choice.

4.3 Tool Impact & Adoption

The Frontier AI Risk Threshold Analyzer has immediate applications:

For AI Labs:

- Self-assessment against all frameworks before deployment
- Gap analysis to identify where internal policies lag competitors
- Compliance documentation for EU AI Act Article 51

For Regulators:

- Objective comparison of lab commitments
- Monitoring framework evolution over time
- Evidence base for harmonization efforts

For Researchers:

- Open dataset for governance research
- Benchmark for evaluating framework effectiveness
- Foundation for developing improved taxonomies

4.4 Limitations

Our work has four key limitations:

Limitation 1: Extraction Errors

While 94% accurate, AI extraction made errors in 6% of cases. Manual review caught these, but fully automated monitoring would propagate mistakes.

Mitigation: We flag low-confidence extractions for human review and maintain version control of all database updates.

Limitation 2: Framework Evolution

Frameworks change frequently (Anthropic RSP is now v2.2, was v1.0 in 2023). Our database represents a snapshot.

Mitigation: We designed the system for easy re-extraction. Re-running scripts takes ~30 minutes and costs ~\$10 in API credits.

Limitation 3: Informal Commitments Missed

Some labs make safety commitments in blog posts, Twitter threads, or informal statements not captured in formal policy documents.

Mitigation: Future work should incorporate web scraping of lab communications and track informal commitments.

Limitation 4: Capability Definitions Are Vague

"CBRN acceleration" and "autonomous replication" lack precise definitions. Different evaluators might assess the same model differently.

Mitigation: Our tool flags when capability definitions are vague and suggests using standardized benchmarks (e.g., METR evals).

4.5 Future Work

We identify four priority areas for extension:

Extension 1: Real-Time Monitoring

Deploy automated web scraping to detect framework changes within 24 hours of publication. Generate alerts when labs update policies.

Extension 2: Standardized Mapping

Work with AI Safety Institutes to develop official $ASL \leftrightarrow CCL \leftrightarrow Preparedness$ equivalence mappings validated by all labs.

Extension 3: Automated Compliance Reports

Generate EU AI Act Article 51 compliance documentation automatically, saving labs 50+ hours of legal review.

4.6 Conclusion

The Frontier AI Risk Threshold Analyzer addresses a critical gap in AI governance infrastructure: the lack of standardized risk definitions across laboratories and regulatory frameworks. By building the first comprehensive, machine-readable database of safety thresholds and deploying an interactive assessment tool, we enable:

- ✓ International coordination through common risk language
- ✓ Regulatory compliance via automated EU AI Act checking
- ✓ Transparency through open-source dataset and methodology
- ✓ Research foundation for developing improved governance frameworks

Our findings reveal significant fragmentation m47 risk tiers with minimal alignment but also demonstrate that AI-powered governance infrastructure can efficiently process and harmonize this complexity.

As frontier AI systems approach ASI-relevant capabilities, the international community needs technical mechanisms for verification, comparison, and enforcement of safety commitments. This tool provides that foundation.

The database, code, and web application are freely available for use by AI labs, regulators, and researchers worldwide. We invite the community to build upon this work toward more robust, harmonized AI governance.

5. References

Anthropic. (2024). Responsible Scaling Policy v2.2. Retrieved from <https://www.anthropic.com/responsible-scaling-policy>

Georgetown Center for Security and Emerging Technology (CSET). (2025). AI Safety under the EU AI Code of Practice. Retrieved from <https://cset.georgetown.edu/>

Google DeepMind. (2024). Frontier Safety Framework v3.0. Retrieved from <https://deepmind.google/safety-framework>

Institute for Law & AI. (2025). The Role of Compute Thresholds for AI Governance. Retrieved from <https://law-ai.org/>

METR (Model Evaluation & Threat Research). (2025). Common Elements of Frontier AI Safety Policies. Retrieved from <https://metr.org/>

OpenAI. (2024). Preparedness Framework Version 2. Retrieved from <https://cdn.openai.com/openai-preparedness-framework-beta.pdf>

European Parliament. (2024). Regulation (EU) 2024/1689 on Artificial Intelligence (AI Act). Official Journal of the European Union.

Epoch AI. (2025). AI Training Compute Trends and Algorithmic Efficiency. Retrieved from <https://epochai.org/>

International AI Safety Institutes Network. (2025). Framework for International Cooperation on AI Safety. Retrieved from <https://aisi.gov/>

Shevlane, T., et al. (2023). Model evaluation for extreme risks. arXiv:2305.15324.

6. Appendix

6.1 Security Considerations

While the Frontier AI Risk Threshold Analyzer is designed as a governance tool, its deployment and use raise several security considerations:

Consideration 1: Information Disclosure Risk

The tool aggregates and exposes framework thresholds in machine-readable format. Malicious actors could use this to identify exactly where safety boundaries lie and optimize harmful models to stay just below detection thresholds.

Mitigation: This information is already public in policy documents. Our tool merely makes it more accessible. The transparency benefit outweighs the marginal risk increase.

Consideration 2: Adversarial Manipulation of Inputs

Users could input false model specifications to generate misleading compliance reports, potentially using these to deceive regulators or investors.

Mitigation: The tool is designed for self-assessment and research, not regulatory certification. All outputs include disclaimers that assessments are preliminary and require expert review.

Consideration 3: AI Extraction Errors

Incorrect threshold extraction could lead labs or regulators to wrong conclusions about compliance status.

Mitigation: We maintain source citations for all extracted data, enabling users to verify against original documents. Regular manual audits catch systematic errors.

Consideration 4: Framework Manipulation

If labs know their frameworks will be analyzed by this tool, they might strategically design policies to appear stricter than they are in practice.

Mitigation: This tool analyzes written commitments. Enforcement monitoring (e.g., are labs actually conducting required evals?) requires separate infrastructure.

6.2 Suggestions for Future Improvements

Technical Improvements:

- Add support for non-English policy documents (Chinese, Korean, French labs)

- Implement differential analysis: "How does v2.3 differ from v2.2?"
- Build API for programmatic access by other governance tools
- Add citation validation: automatically check if referenced evals exist

Policy Improvements:

- Collaborate with AI Safety Institutes to create official mappings
- Develop standardized capability evaluation protocols
- Create "Framework Health Score" measuring clarity and enforceability
- Design treaty-ready risk tier definitions

Deployment Improvements:

- Migrate to dedicated infrastructure for 99.99% uptime
- Add user authentication for tracking institutional usage
- Implement feedback system for labs to suggest corrections
- Create email alerts for framework changes

6.3 Dataset Access

The complete extracted database is available at the github repo

Files include:

- frameworks.json: All 12 lab frameworks (47 tiers)
- eu_compliance.json: EU AI Act requirements
- compute_thresholds.json: FLOP threshold analysis

Format: JSON (machine-readable)

License: MIT (free for any use)

Size: ~500KB

Last updated: February 2026

6.4 Acknowledgments

This research was conducted at the Technical AI Governance Challenge 2026, organized by Apart Research. We thank:

- Apart Research for hackathon organization and technical infrastructure
- OpenAI for API credits enabling AI-powered extraction
- METR, Georgetown CSET, and Institute for Law & AI for foundational research
- All AI labs for publishing transparent safety frameworks
- The AI safety community for feedback and validation

Special thanks to the AI Safety UAE community for ongoing support and collaboration.