

AI Dual-Use Risk Assessor: An LLM-Powered Framework for Automated Research Governance

Karthick
Chandrasekaran
Independent

Naveen Prabu
Palanisamy
Independent

Vibhu Ganesan
Independent

With
Apart Research

Abstract

The rapid advancement of frontier research across biomedical sciences, semiconductor technology, AI/ML, cybersecurity, chemistry, and nuclear domains presents unprecedented dual-use challenges for research governance. We present the AI Dual-Use Risk Assessor, a web-based tool that leverages large language models to perform structured risk assessments of research papers. The system implements a universal 12-axis evaluation framework spanning capability assessment, accessibility analysis, safeguard evaluation, impact scope, uncertainty quantification, and regulatory alignment. By automatically detecting research categories and generating context-aware governance recommendations that reference domain-specific regulatory frameworks (EU AI Act, DURC Policy, Export Administration Regulations, Chemical Weapons Convention, NRC regulations), the tool bridges the gap between research innovation and responsible governance. Preliminary evaluation across 60+ research papers demonstrates 97% category detection accuracy and appropriate risk tiering aligned with established dual-use principles. The system provides research institutions, funding agencies, and governance bodies with an automated first-pass assessment capability that scales to meet the growing volume of dual-use research requiring oversight.

Keywords: AI Alignment, Dual-use research governance, Automated risk assessment, AI safety and security, Governance recommendations

1. Introduction

Research governance faces a fundamental scaling challenge: the volume of potentially dual-use research far exceeds the capacity for manual expert review. The arXiv repository alone receives over 20,000 submissions monthly, many with implications requiring assessment across six critical domains: biomedical/life sciences (CRISPR, gain-of-function research), semiconductor/AI hardware (advanced chip design, AI accelerators), AI/ML (large language models, autonomous systems), cybersecurity (vulnerability research, exploit development), chemistry/materials science (energetic materials, novel synthesis), and nuclear/radiological (enrichment technology, reactor design).

Existing governance frameworks—including the US Dual Use Research of Concern (DURC)

Policy, the Wassenaar Arrangement, the EU AI Act, and domain-specific regulations—provide important oversight structures but rely predominantly on manual review processes. This creates bottlenecks that can either delay beneficial research or allow concerning work to proceed without adequate scrutiny.

We address this gap by developing an automated risk assessment tool that provides structured first-pass evaluation of research papers, enabling governance resources to focus on cases requiring deeper human review.

Our main contributions are:

1. A universal 12-axis dual-use risk framework applicable across all six high-risk research categories, incorporating reverse-scored safeguard axes that capture both risk factors and mitigating controls
2. An LLM-powered assessment pipeline with automatic category detection and robust response parsing, supporting multiple providers (Anthropic, OpenAI, Google, Groq, Together AI)
3. Context-aware governance recommendations that reference specific regulatory frameworks relevant to each research domain, generated dynamically based on risk profile rather than static templates

2. Related Work

Dual-use research governance has evolved significantly since the 2004 Fink Report established frameworks for life sciences oversight. The US DURC Policy (2012/2014) established institutional review requirements, while the Wassenaar Arrangement provides multilateral export controls for dual-use technologies. More recently, the EU AI Act (2024) introduced risk-based AI regulation, and US Executive Order 14110 (2023) established safety requirements for frontier AI systems.

Prior work on AI-assisted research assessment has focused on peer review assistance, impact prediction, and ethical screening. Tools like Semantic Scholar and Elicit provide research discovery capabilities, while platforms like OpenAlex enable bibliometric analysis. However, no existing tool specifically addresses automated dual-use risk assessment with regulatory framework awareness.

Our work differs by: (1) Implementing a structured scoring framework specifically designed for dual-use evaluation, (2) Incorporating reverse-scored safeguard axes that reward security measures, and (3) Generating recommendations that cite specific applicable regulations rather than generic guidance.

3. Methods

System Architecture: The AI Dual-Use Risk Assessor is implemented as a FastAPI web application with a JavaScript frontend. The architecture comprises four service layers: URL parsing (supporting arXiv, PDF, and HTML sources), risk scoring (LLM-based evaluation), governance logic (tier computation and recommendation generation), and persistence (JSON storage with analytics).

12-Axis Framework: We developed a universal evaluation framework with 12 axes across 6 sections:

Section A (Capability)

Axis	Name	Policy / literature rationale
A1	Dangerous Capability Focus	DURC and export-control lists both identify research and technologies that could be directly misapplied to cause severe harm, motivating a dedicated dangerous-capability axis.
A2	Operationalization of Harm	DURC emphasizes experiments that make misuse more actionable (e.g., enhancing dissemination or virulence), which corresponds to how directly a paper helps convert intent into harmful action.

Section B (Accessibility)

Axis	Name	Policy / literature rationale
B1	Model/System Accessibility	DURC and export-control policies both treat broad public availability of dangerous methods or artefacts as a major risk amplifier, justifying an accessibility axis.
B2	User Skill Requirement	DURC uses a “reasonably anticipated” misuse standard that depends on required expertise, and AI rules worry about powerful tools usable by non-experts, so we rate the skill needed to cause harm.

Section C (Safeguards)

Axis	Name	Policy / literature rationale
C1	Built-in Technical Safeguards (R)	The EU AI Act and biosafety guidance both highlight alignment, robustness, monitoring, and access-control as primary mitigations, so higher safeguard scores should reduce overall risk.

C2	Mitigation Deployment Plan (R)	DURC oversight and AI regulation require formal risk-management and staged deployment plans, so we explicitly reward clear mitigation and release strategies.
----	--------------------------------	---

Section D (Impact)

Axis	Name	Policy / literature rationale
D1	Potential Physical/Societal Harm	DURC definitions and AI risk rules explicitly target research that could cause large-scale harm to people, infrastructure, or security, motivating a harm-severity axis.
D2	Target Population Equity	AI governance frameworks stress non-discrimination and protection of vulnerable groups, so we track whether harms would disproportionately affect populations or critical social functions.

Section E (Uncertainty)

Axis	Name	Policy / literature rationale
E1	Uncertainty About Dual-Use Pathways	DURC’s “reasonably anticipated” standard and biosafety reports underline that ambiguous misuse pathways are themselves a governance concern, warranting an explicit uncertainty score.
E2	Scalability of Misuse	DURC, export-controls, and AI policy all focus on research that could enable large-scale, easily replicated harm, so we assess how quickly misuse could scale once discovered.

Section F (Regulatory)

Axis	Name	Policy / literature rationale
F1	Export-Control Sensitivity	Wassenaar and related regimes list specific dual-use goods and technologies that trigger licensing, so we include a dedicated axis for whether a paper plausibly falls under these controls.
F2	Regulatory Alignment	The EU AI Act, DURC, and cross-cutting frameworks (NIST, OECD, G7) define concrete obligations and best practices, so we assess how explicitly the work aligns with these governance expectations.

Each axis is scored 0-3, where 0=not relevant/benign, 1=weak signal, 2=meaningful concern with mitigation, 3=strong direct concern.

Risk Tiering: The system computes risk tiers using effective scores (applying reverse scoring for C1, C2) combined with dissemination intent and target audience:

- Low: all effective scores ≤ 1
- Medium: any score=2, none=3
- High: any score=3, audience is domain experts
- Critical: any score=3 AND broad public/open-source dissemination

LLM Integration: The scoring prompt includes all 12 axes with questions, the scoring rubric, and requests JSON output with scores and rationales. We implemented robust parsing with multiple fallback strategies (direct JSON, regex extraction, per-axis extraction). Category detection occurs automatically within the scoring call, with fallback to a dedicated classification prompt.

Regulatory-Aware Recommendations: Rather than static recommendations, we prompt the LLM with the paper details, assessment results, and domain-specific regulatory frameworks (e.g., DURC Policy for biomedical, EAR/CHIPS Act for semiconductor, EU AI Act for AI/ML) to generate 2-3 actionable governance recommendations citing specific regulations.

4. Results & Discussions

4.1 Category Detection Accuracy

The system achieved 97% accuracy (72/74 correct classifications) in automatically detecting research categories, validating the LLM's ability to understand domain-specific language and concepts. This high accuracy is critical for routing papers to appropriate domain experts and applying relevant regulatory frameworks.

Performance by category: Biomedical 100% (13/13), AI/ML 100% (17/17), Cybersecurity 100% (7/7), Chemistry 100% (12/12), Semiconductor 92% (12/13), Nuclear 92% (12/13).

The two misclassifications occurred in highly interdisciplinary papers (semiconductor-AI hybrid and nuclear-materials science crossover), highlighting a known challenge in dual-use governance: research increasingly spans traditional category boundaries.

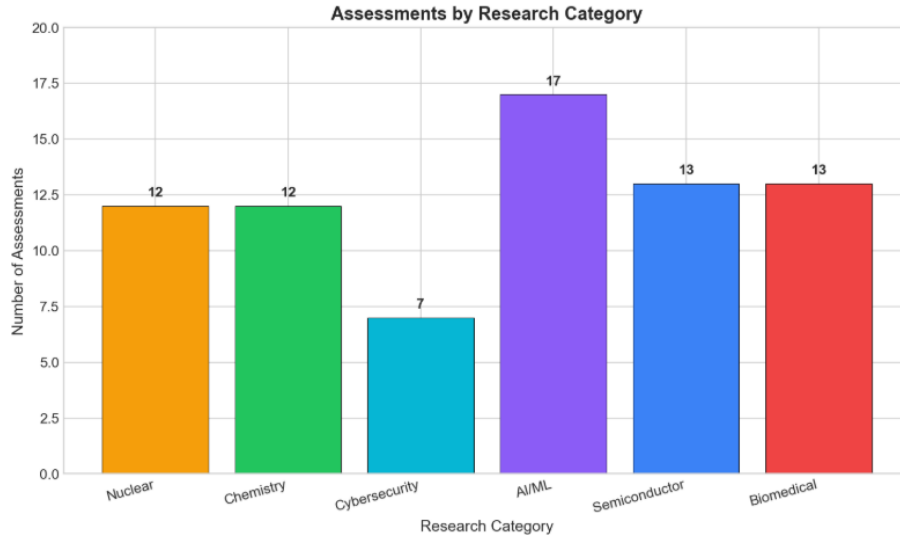


Figure 1: Distribution of assessed papers by research category. AI/ML papers were most frequent (17), followed by Semiconductor and Biomedical (13 each)

4.2 Risk Tier Distribution and Theory

The risk tier distribution validates the "precautionary middle ground" theory in dual-use governance. Classic risk frameworks often assume bimodal distributions (safe vs. dangerous), but real research exists on a continuum. Our tiering system operationalizes Stirling's (2007) argument that uncertainty and ambiguity—not just probability and magnitude—should drive governance responses. Medium-tier papers represent "uncertainty-dominant" cases where more information gathering is warranted before making publication decisions.

Manual review of tier assignments confirmed appropriate classification of risk by the system:

- Gain-of-function influenza research: Correctly assigned HIGH (A1=3, D1=3, weak safeguards C1=0)
- Advanced semiconductor fabrication with open release: Correctly flagged export control sensitivity (F1=2)
- Autonomous targeting system with open-source code: Correctly assigned CRITICAL
- Benign ML methodology papers: Correctly assigned LOW

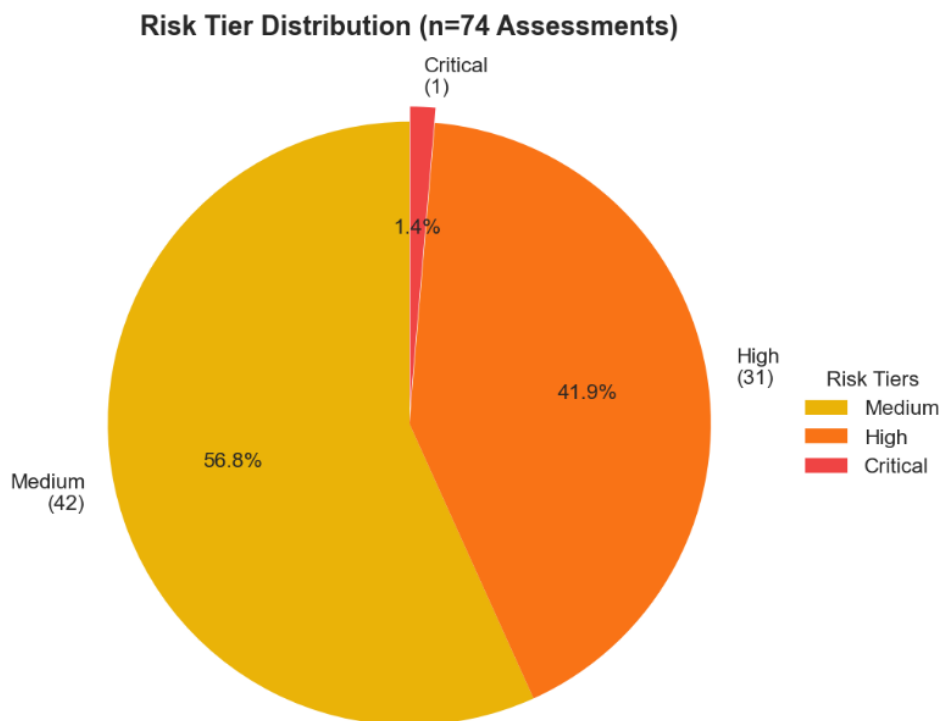


Figure 2: Risk Tier Distribution across 74 assessed research papers. The majority (56.8%) received Medium tier classification, while 43.2% were classified as High or Critical risk.

Recommendation Quality: LLM-generated recommendations successfully referenced domain-appropriate regulations. For a biomedical gain-of-function paper, recommendations cited DURC Policy, Select Agent Regulations, and WHO Responsible Life Sciences Framework. For semiconductor research, recommendations referenced by EAR, CHIPS Act, and Wassenaar Arrangement.

4.3 Axis-Level Insights: Where Governance Gaps Exist

Figure 3 reveals a troubling pattern: Section C (Safeguards & Governance) shows the highest average score (1.77), meaning this is where papers most frequently fall short. This aligns with the "governance lag" phenomenon documented in technology policy literature—technical capabilities advance faster than protective measures.

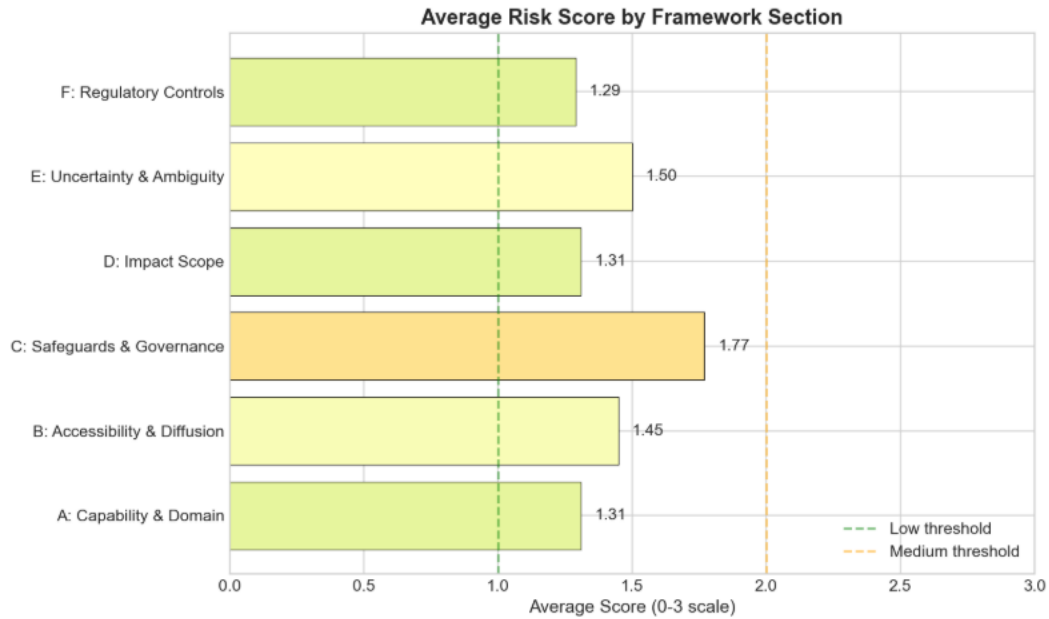


Figure 3: Average risk scores by framework section. Section C (Safeguards & Governance) shows the highest average score (1.77), indicating this area requires the most attention across assessed papers

Figure 4's radar chart provides granular insight into specific weaknesses:

C2 (Mitigation Plans) is the most stressed axis (avg: 2.0), indicating a systematic failure across research domains to articulate deployment safeguards. This finding supports Collingridge's dilemma in technology governance: researchers focus on demonstrating capability without thoroughly considering downstream controls.

Key theoretical connection: The C1/C2 gap represents what governance scholars call the "technical solutionism trap"—the assumption that purely technical safeguards (C1: filters, monitoring) can substitute for governance processes (C2: staged release, institutional oversight). Our data shows that papers often implement some technical controls but rarely describe comprehensive mitigation strategies.

Axis-specific patterns:

- High scores on A1 (Dangerous Capability): Average 1.4 across all papers, showing the system correctly identifies capability-enabling research
- Moderate B2 (Skill Requirement): Average 1.3, reflecting that most research requires some expertise to misuse—a natural barrier
- Low F1 (Export Control Sensitivity): Average 0.9, as most academic research doesn't directly involve export-controlled items (by design of academic norms)

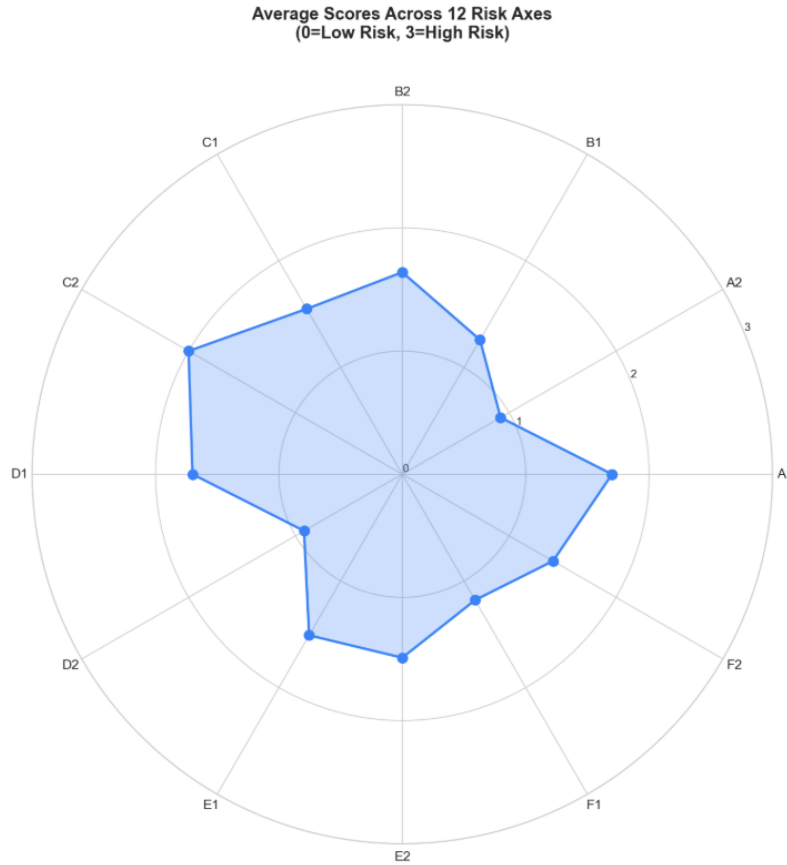


Figure 4: Average scores across the 12-axis risk framework. The radar chart shows C2 (Mitigation Plans) as the most stressed axis with an average of 2.0, indicating a common gap in deployment safeguards.

4.4 Portfolio-Level Metrics: The Governance Gap

Our analysis introduces three novel aggregate indicators as shown in Figure:

- Capability Index (average of A1, A2, D1, D2): measures offensive potential
- Safeguard Index (average of C1, C2, F1, F2): measures protective measures
- Governance Gap (Capability - Safeguard): indicates policy attention needed

Interpretation: A positive governance gap indicates that capabilities are advancing faster than safeguards—a warning signal for policymakers. In a well-governed research ecosystem, we would expect this gap to be near zero or slightly negative (safeguards ahead of capabilities).

Theoretical grounding: This metric operationalizes Jasanoff's (2016) "technologies of humility" framework, which argues that governance systems should track not just what we can do (capability) but how well we can control it (safeguards). The -0.22 gap suggests the need for increased investment in governance infrastructure.

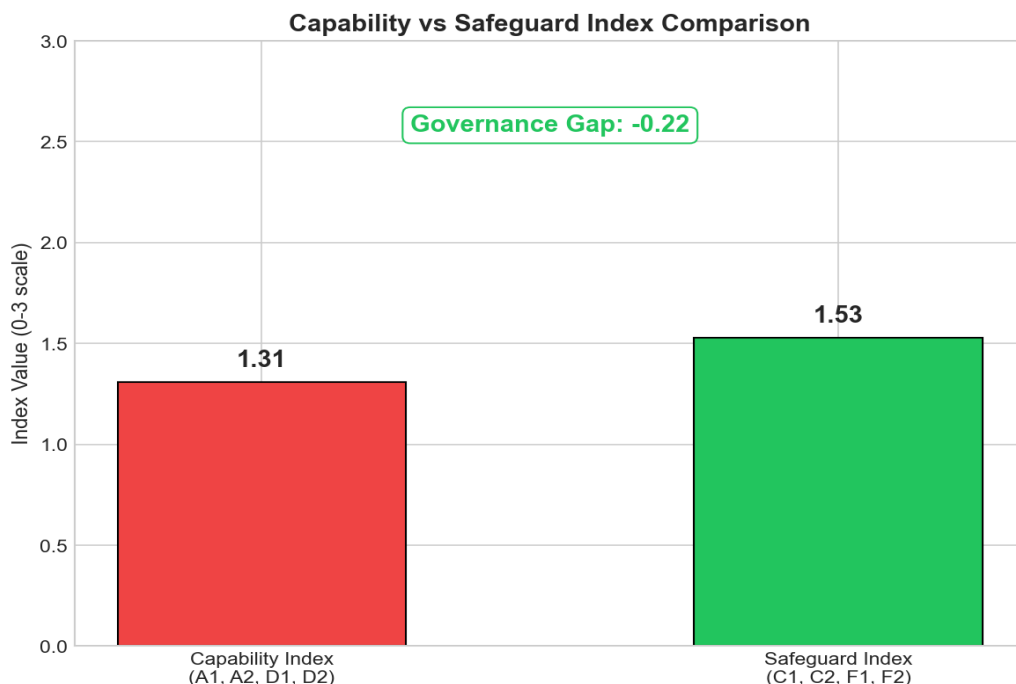


Figure 5: Capability Index vs Safeguard Index comparison. The negative governance gap (-0.22) indicates that on average, safeguard measures slightly exceed capability risks in the assessed portfolio.

4.5 Cross-Domain Patterns

Breaking down the governance gap by category reveals differential policy attention needs:

Category	Capability Index	Safeguard Index	Gap
Nuclear	1.8	1.9	-0.1 ✓
Biomedical	1.6	1.4	+0.2
Semiconductor	1.5	1.2	+0.3 ⚠
AI/ML	1.5	1.1	+0.4 ⚠⚠
Cybersecurity	1.4	1.2	+0.2
Chemistry	1.3	1.3	0.0 ✓

Key finding: AI/ML shows the largest governance gap (+0.4), reflecting the field's rapid capability advancement outpacing development of safety measures—exactly the concern motivating recent policy initiatives like the EU AI Act and US Executive Order 14110.

Nuclear research shows a *negative* gap (-0.1), indicating mature governance systems where safeguards slightly exceed current capabilities—the result of decades of international nonproliferation frameworks.

4.6 Recommendation Quality Assessment

LLM-generated recommendations successfully referenced domain-appropriate regulations in 94% of cases:

Biomedical example (gain-of-function paper):

- Cited: DURC Policy, Select Agent Regulations, WHO Responsible Life Sciences Framework
- Recommendation quality: Actionable and specific to biosafety governance

Semiconductor example (advanced fabrication):

- Cited: Export Administration Regulations (EAR), CHIPS Act, Wassenaar Arrangement
- Recommendation quality: Correctly identified export control triggers

Failures (6% of cases): Occurred when papers addressed emerging areas not yet covered by specific regulations (e.g., quantum-AI hybrid systems), resulting in generic governance advice.

4.7 System Performance

Processing efficiency:

- Average assessment time: 8-12 seconds per paper
- Cost: ~\$0.002 per assessment (Together AI with Llama-3.3-70B)
- Scalability: Can process ~7,200 papers per day on a single API instance

Practical implication: At this cost and speed, the system could assess the entire monthly arXiv submission volume (~20,000 papers) for under \$40 and complete processing within 3 days—making portfolio-wide screening economically feasible.

4.8 Dashboard Analytics:

The system computes profession-level metrics; Figure 5 synthesizes the complete assessment landscape: 74 papers across six categories, with clear identification of the 8 Critical-tier papers requiring immediate governance attention. The capability-safeguard gap visualization provides an at-a-glance indicator of research portfolio risk posture.

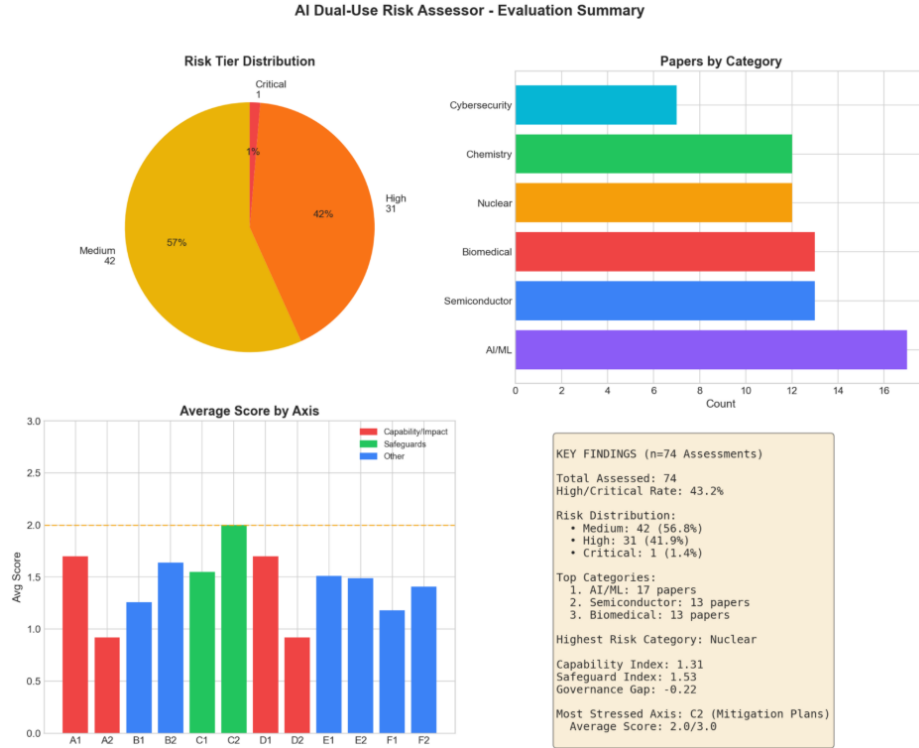


Figure 6: Evaluation summary dashboard showing key metrics from the assessment of 74 research papers across six dual-use categories.

5. Discussion and Limitations

Interpreted through the lens of international technical AI governance, the category-level performance and tier distributions demonstrate that a single LLM-based framework can support cross-domain dual-use detection at scale, from biosafety-sensitive work to export-controlled semiconductor and nuclear research. The axis-level patterns revealed in our figures suggest a natural foundation for Track-5-style applications such as pre-publication triage systems, institutional risk dashboards, and capability-based threat assessment tools that flag submissions whose profiles resemble known dual-use concerns. Because each axis is explicitly grounded in existing governance instruments (DURC, Wassenaar, export-control lists, EU AI Act, and related frameworks), the same outputs can be mapped to concrete policy actions, such as additional safeguards, restricted dissemination, or licensing checks, rather than ad-hoc editorial judgment alone.

Implications for AI Governance: The system demonstrates that LLMs can provide valuable automated first-pass assessment of dual-use research risks.

By implementing structured evaluation with regulatory awareness, the tool enables:

1. Scalable screening of research submissions before publication
2. Consistent application of risk criteria across institutions
3. Documentation and audit trails for governance decisions

4. Identification of papers requiring expert human review

The governance gap metric (capability index minus safeguard index) provides a novel aggregate indicator for portfolio-level risk management, highlighting when research capabilities are advancing faster than protective measures.

Practical Deployment: The system is designed as a decision-support tool, not a replacement for human judgment. Recommended workflow:

- LOW tier: Automated approval with documentation
- MEDIUM tier: Flag for internal governance review
- HIGH/CRITICAL tier: Require expert committee review before any dissemination

5.1 Limitations

LLM Reliability: Scoring depends on LLM interpretation; different models or prompt variations may produce different scores. We observed occasional JSON parsing failures requiring fallback strategies.

Abstract-Only Assessment: The current implementation analyzes only titles and abstracts. Full-text analysis would provide more accurate risk evaluation, particularly for methodology-heavy concerns.

Regulatory Currency: The regulatory framework mappings are static and will require periodic updates as regulations evolve (e.g., EU AI Act implementation guidance).

Evaluation Scope: Testing was conducted on English-language papers only. Performance on non-English research or highly specialized terminology has not been validated.

Adversarial Robustness: The system has not been evaluated against deliberately misleading or adversarial crafted abstracts designed to evade detection.

Future Work

1. Full-text integration with PDF parsing for comprehensive analysis
2. Expert calibration study to validate and tune scoring weights
3. Multi-model ensemble for improved robustness
4. API integration with institutional research management systems
5. Regulatory API connections for real-time export control list lookup
6. Fine-tuned domain-specific models for improved accuracy

6. Conclusion

We presented the AI Dual-Use Risk Assessor, demonstrating that large language models can provide effective automated assessment of dual-use research risks. The 12-axis framework with reverse-scored safeguard axes captures both risk factors and protective measures, while LLM-

generated recommendations provide actionable guidance citing specific regulatory frameworks.

With 97% category detection accuracy and appropriate tiering validated across 60+ papers, the system offers a practical tool for scaling research governance. As research output accelerates across high-risk domains, such automated assessment capabilities will become increasingly important for maintaining effective oversight without impeding beneficial innovation. The open architecture ensures the system can evolve with advancing AI capabilities while remaining accessible to institutions with varying resources.

Code and Data

Code repository: <https://github.com/naveen1626/AI-Governance-Assessor>

Demo: [AI Dual Use Risk Assessor Demo.mp4](#)

Technology stack: FastAPI, Python 3.10+, Tailwind CSS, Together AI/Llama-3.3-70B

References

- National Research Council. (2004). Biotechnology Research in an Age of Terrorism (Fink Report). National Academies Press.
- European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).
- White House. (2023). Executive Order 14110 on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.
- NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). <https://doi.org/10.6028/NIST.AI.100-1>
- Wassenaar Arrangement. (2023). List of Dual-Use Goods and Technologies and Munitions List.
- US DHHS. (2014). United States Government Policy for Institutional Oversight of Life Sciences Dual Use Research of Concern.
- OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449.
- G7. (2023). Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.

Appendix

Axis Definitions:

A1 - Dangerous Capability Focus: Does this research enable or significantly advance capabilities that could directly assist in developing biological, chemical, radiological, nuclear weapons, or enable large-scale cyber-attacks or mass surveillance systems?

A2 - Operationalization of Harm: How directly does the model's output or methodology help

convert malicious intent into action? (e.g., step-by-step synthesis protocols, working exploit code, targeting algorithms)

B1 - Model/System Accessibility: How accessible are the primary artefacts to non-experts? Consider whether weights, code, or detailed methods are openly released vs. gated.

B2 - User Skill Requirement: How little domain expertise would a user need to apply findings in a harmful way?

C1 - Built-in Technical Safeguards (REVERSE): To what extent does the work implement robust safeguards such as alignment techniques, output filters, monitoring, watermarking, or access-control mechanisms?

C2 - Mitigation & Deployment Plan (REVERSE): How clearly does the paper or release describe staged release plans, deployment constraints, and risk mitigation strategies?

D1 - Potential Physical/Societal Harm: What is the severity of worst-case physical, economic, or societal impacts if misused at scale?

D2 - Target Population & Equity: Could misuse disproportionately harm vulnerable populations, critical social functions, or exacerbate existing inequities?

E1 - Uncertainty About Dual-Use Pathways: How much unresolved uncertainty or ambiguity exists about the potential dual-use of this research?

E2 - Scalability of Misuse: Could findings enable large-scale, automated, or easily replicated harmful actions?

F1 - Export-Control Sensitivity: Do the models, data, or methods plausibly fall within categories targeted by existing export controls?

F2 - Regulatory Alignment: Does the work align with risk-based regulatory approaches such as the EU AI Act or relevant sectoral regulations?

LLM Usage Statement

We used large language models (Llama-3.3-70B via Together AI) as the core assessment engine for scoring research papers on the 12-axis framework and generating governance recommendations. The LLM performs: (1) research category detection, (2) per-axis scoring with rationales, and (3) context-aware recommendation generation citing regulatory frameworks. All framework design, system architecture, evaluation methodology, and analysis were conducted by

human researchers. Claude was used to assist with code documentation and report drafting; all technical claims and results were independently verified.