Safety First! Modelling Requirements from GPT-5 System Card using Lightweight Safety Models

MoDRE '26, Montreal

Kalvin Thuan-Phong Khuu – McMaster University

Nicolas Lacroix – Université Côte d'Azur

Mireille Blay-Fornarino – Université Côte d'Azur

Sébastien Mosser – McMaster University

Centre for Software Certification (McSCert)

McMaster University

UNIVERSITÉ CÔTE D'AZUR

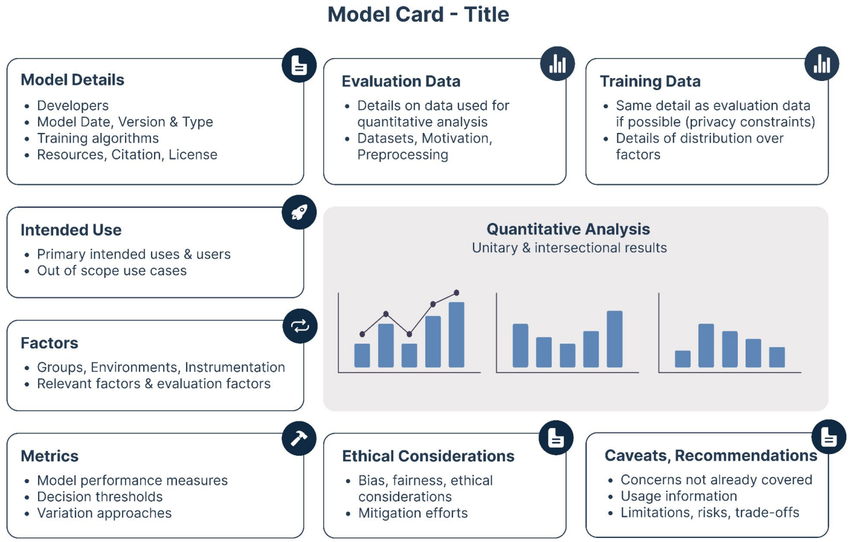
i3S sophia antipolis

Inria

Core Problem

Model/System Cards

- Model/System cards are the main AI model and the system surround it documentation
- Capabilities, Limitations, Safety Properties/Mitigations
  - OpenAI, Anthropic, Google
- However,
  - Unstructured, lengthy documentation written in natural language
- As such,
  - What requirements in system cards are actually being addressed?
  - Are the claims backed by adequate evidence?
  - Are system cards traceable for regulations?



*We study the GPT-5 system card to understand what **requirements** are elicited by OpenAI for their flagship system, and how **lightweight safety model** can contribute to their further analysis*

Case Study

Motivation

- A lawsuit filed Aug. 2025 against OpenAI in California for the wrongful death of Adam Raine
  - GPT-4 was used as Raine's personal diary
- The OpenAI company acknowledged that "*there have been moments where our systems did not behave as intended in sensitive situations*".
- → Need for properly defining requirements to address risks related to the use of ML-based systems

Parents of teenager who took his own life sue OpenAI

27 August 2025

Share Save

Nadine Yousif

BBC News

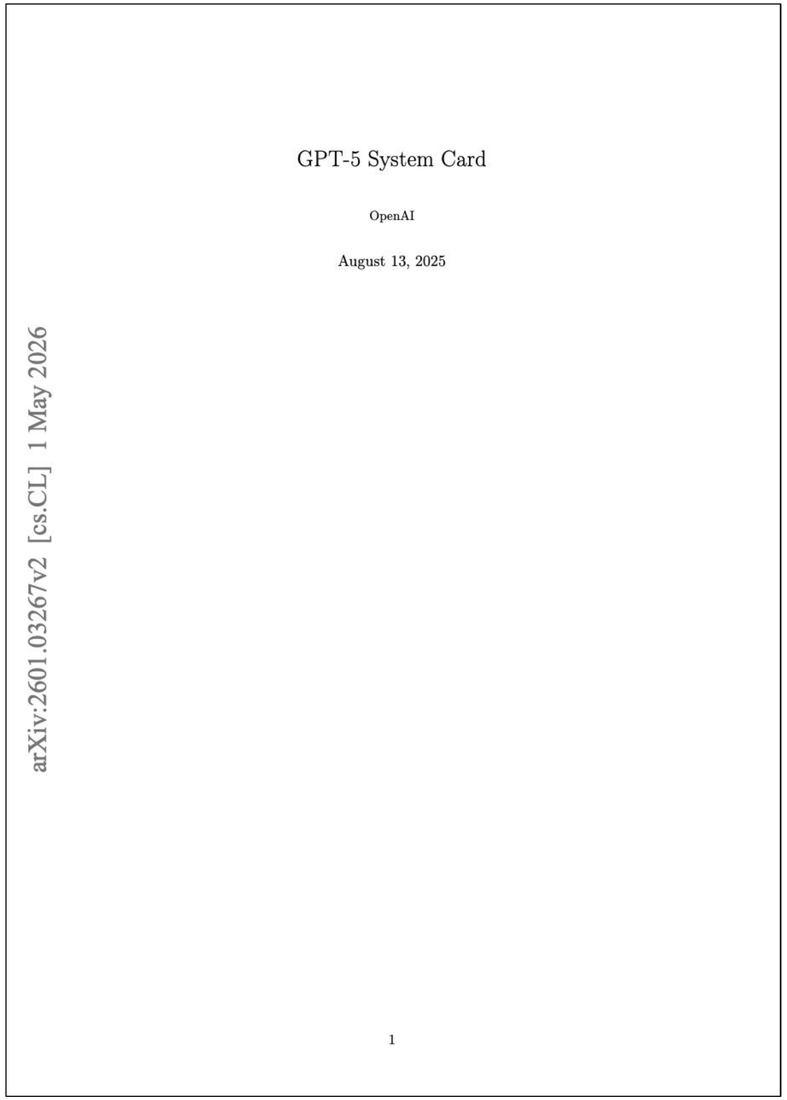
The Raine Family

A California couple are suing OpenAI over the death of their teenage son, alleging its

Case Study

GPT-5 system card

- Manual analysis (Extraction and Modelling) on GPT-5 system card
  - Published in August 2025 by OpenAI
  - 60 pages, covering *gpt-5-thinking* and *gpt-5-main*
  - *Latest system card from the GPT family
    - Over a dozen cards since 2020
  - Progressive adoption of this documentation
    - 1 page for GPT-3 (2020) to 60 pages for GPT-5 (2025)

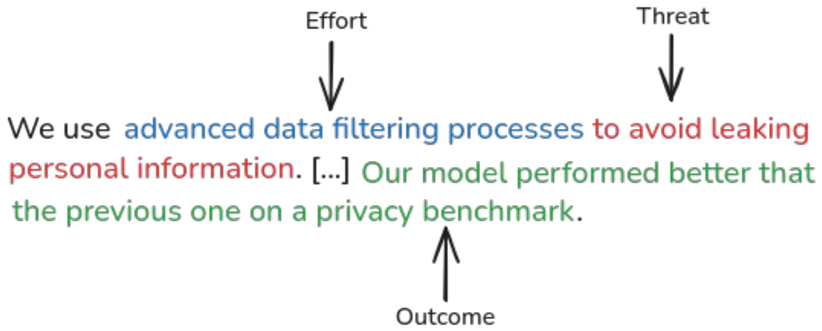


Approach

Methodology

Claim collection:

- 2 annotators (AI safety & ML expert) + 1 mediator (RE/safety expert)
- Claim on capabilities, limitations, and/or mitigated threats
  - Consist of a threat, effort, and outcome



Requirements elicitation:

- Bottom-Up approach to derive implicit requirements
- Clarification of ambiguous terms
  - Ex.: “deception”, “jailbreak”, “sycophantic behavior”
- Classification of requirements using **MIT AI Risk Taxonomy**
  - *Discrimination & Toxicity; Privacy & Security; Misinformation; Malicious Actors & Misuse; Human-Computer Interaction; Socioeconomic & Environmental; AI System Safety, Failures & Limitations*

Key findings

Requirements found

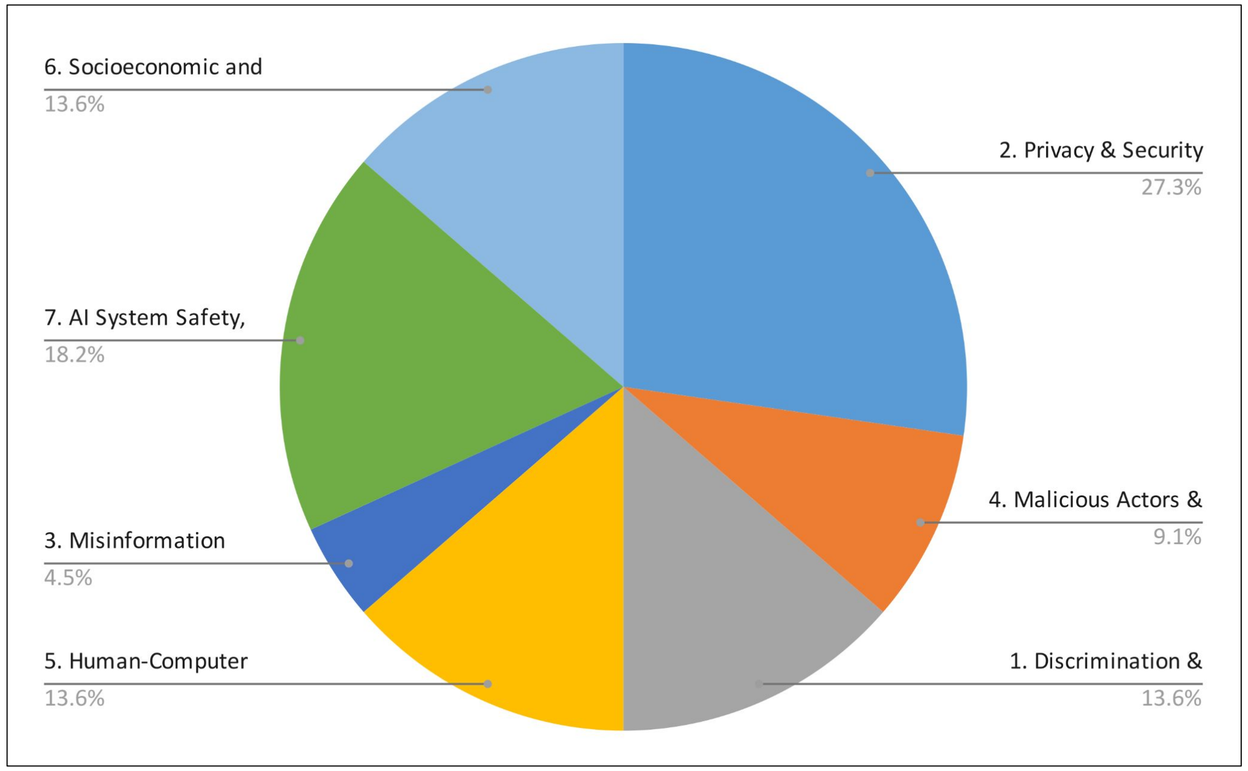


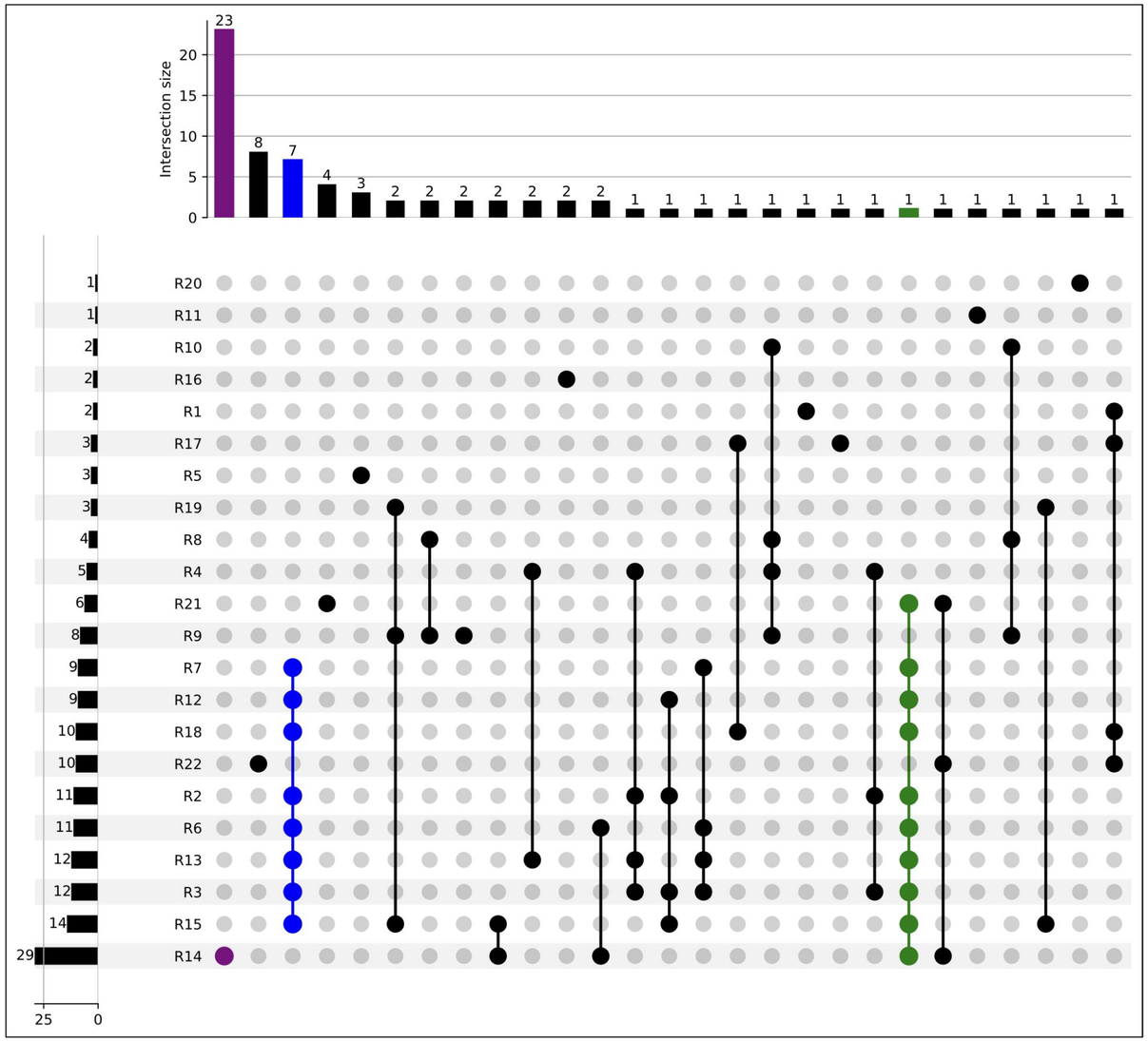
TABLE III
REQUIREMENTS ELICITED FROM GPT-5 SYSTEM-CARD, CLASSIFIED USING THE MIT AI RISK DOMAIN TAXONOMY.

Phase	Risk Domain	Risk Subdomain	ID	Requirement
Pre	2. Privacy & Security	2.1: Compromise of privacy by leaking or correctly inferring sensitive information	R1	GPT-5 shall use training datasets that has been ethically scraped with information from the Internet
			R17	GPT-5 shall be trained on data that have passed a quality assessment process
Post	1. Discrimination & Toxicity	1.2: Exposure to toxic content	R3	GPT-5 shall minimize the generation of disallowed content in accordance with their policy
		1.3: Unequal performance across groups	R11	GPT-5 shall support multilingual natural language processing and generation
			R20	GPT-5 shall be fair
	2. Privacy & Security	2.2: AI system security vulnerabilities and attacks	R6	GPT-5 shall be more or on par resilient to jailbreak attacks in comparison to previous models
			R7	GPT-5 shall perform better in prompt injection evaluations in comparison to previous models
			R15	GPT-5 shall incorporate a proper safeguard architecture, both at a model and at a system level
	3. Misinformation	3.1: False or misleading information	R8	GPT-5 shall produce hallucinations at a lower rate in comparison to previous models
	4. Malicious Actors & Misuse	4.0: Malicious use	R2	GPT-5 shall correctly classify user messages as a range of safe to unsafe
		4.2: Cyberattacks, weapon development or use, and mass harm	R21	GPT-5 shall reject any operational uplift for cybersecurity attacks
	5. Human-Computer Interaction	5.1: Overreliance and unsafe use	R5	GPT-5 shall include ongoing research and evaluation of user interactions that may indicate emotional dependency, unhealthy attachment, or any other forms of psychological distress
			R10	GPT-5 shall outperform previous models on health-related benchmarks results
		5.2: Loss of human agency and autonomy	R4	GPT-5 shall reduce sycophancy response behaviors in comparison to previous models
	6. Socioeconomic and Environmental	6.1: Power centralization and unfair distribution of benefits	R16	GPT-5 shall be released to the public, including all other models from the same generation
		6.2: Increased inequality and decline in employment quality	R22	GPT-5 shall maximize its self-improvement capabilities within a defined scope
		6.5: Governance failure	R19	GPT-5 shall be subject to continuous monitoring
	7. AI System Safety, Failures, & Limitations	7.0: AI system safety, failures, & limitations	R13	GPT-5 shall demonstrate safer performance than previous models in terms of different safety domains (Frontier harms, content safety, psychosocial harms)
		7.1: AI pursuing its own goals in conflict with human goals or values	R9	GPT-5 shall demonstrate deceptive behaviour at a lower rate in comparison to previous models
		7.2: AI possessing dangerous capabilities	R12	GPT-5 shall reject any requests related to the planning or execution of violent attack with better performance to previous models
			R14	**GPT-5 shall exhibit little to no chemical or biological (CB) capabilities with similar or lower performance than previous models**
Other	2. Privacy & Security	2.1: Compromise of privacy by leaking or correctly inferring sensitive information	R18	GPT-5 shall prevent the generation of outputs that would expose, reconstruct, or infer personally identifiable information

Key findings

Analysis

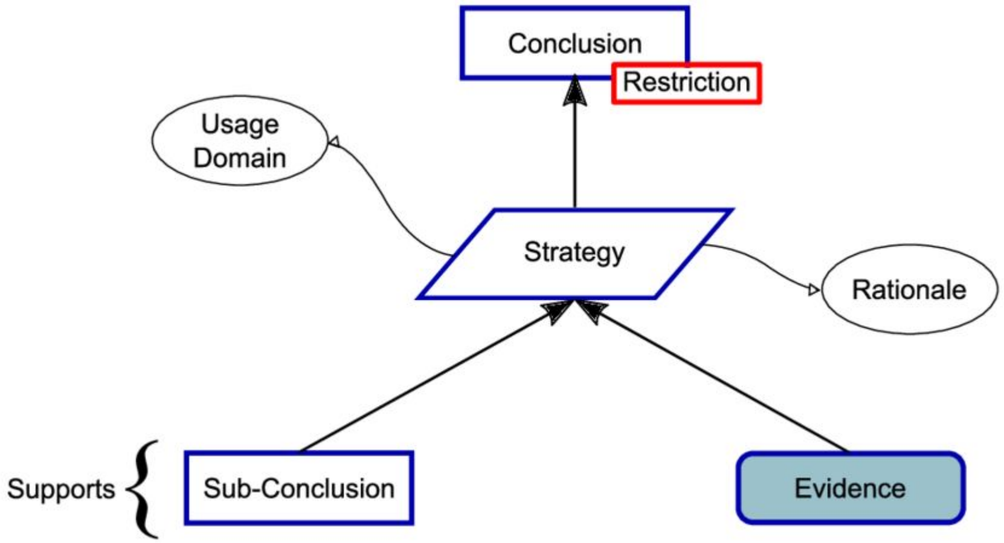
- **74 claims** linked to **22 requirements**
- Requirements are unevenly covered:
  - R14 (chemical/biological capability) supported by 29 claims (~38% of all claims)
  - Meanwhile, R11 (multilingual) & R20 (fairness) is supported by a single evidence
    - With R20 solely relying on the BBQ benchmark without any definition for “fairness”
- Requirements are tightly coupled:
  - 8 requirements (R2, R3, R6, R7, R12, R13, R15, R18) share substantial common evidence



Modelling Requirement

Background

- **Justification Diagrams (JDs) :** Lighter than GSN, still expressive, industrial scale in safety-critical systems
  - Conclusion, Strategy, Sub-Conclusion, Evidence
- R14: *GPT-5 shall exhibit little to no chemical or biological (CB) capabilities with similar or lower performance than previous models*

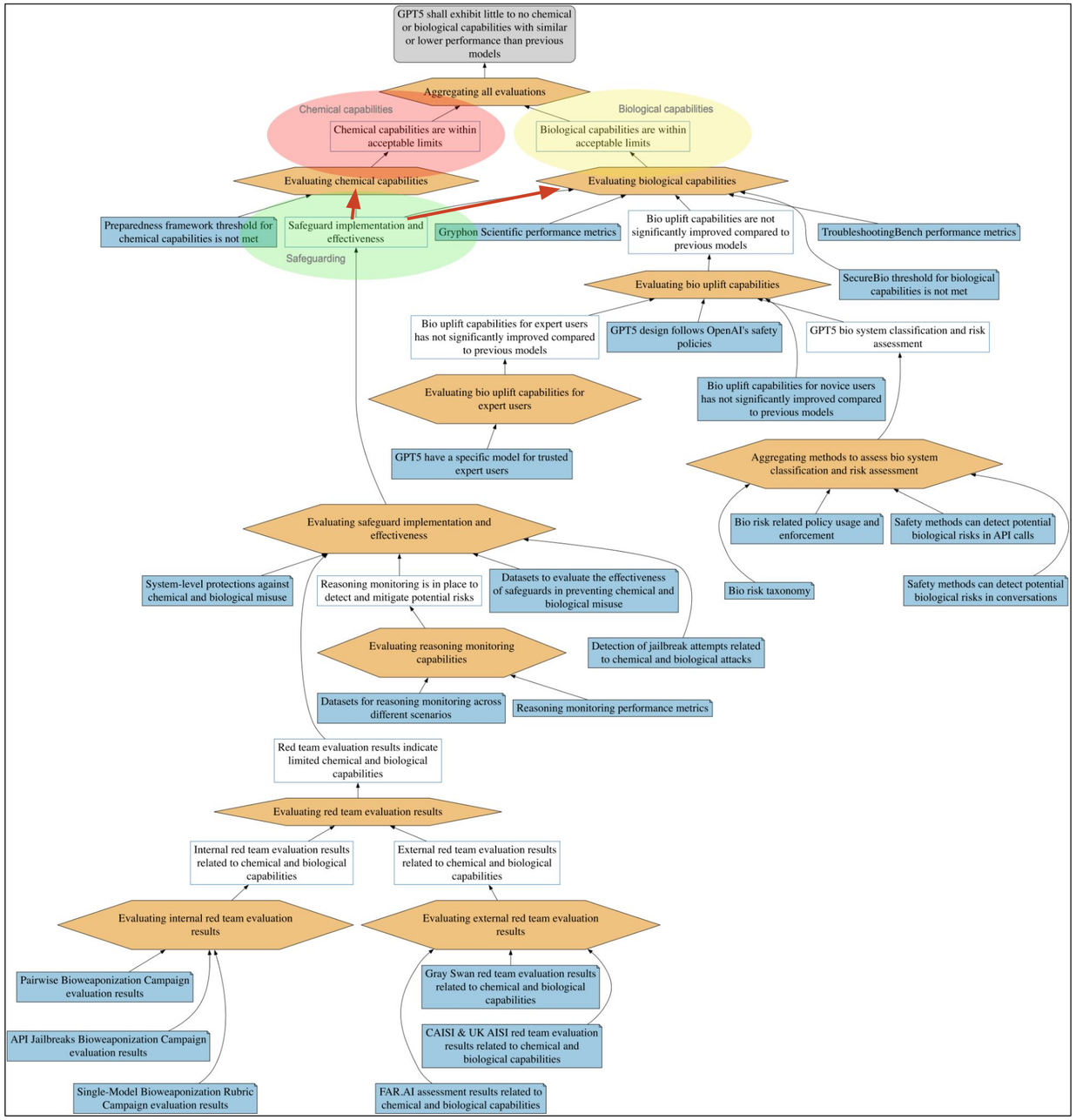


JPIPE
EST. 2023

Modelling Requirement

Modelling R14

- **Asymmetric treatment**
  - Chemical capability relies on a single threshold assessment with safeguards
  - Biological capability is justified with performance metrics and multiple biological uplift mitigation strategies
- **Heavy reliance on safeguarding**
  - Safeguarding accounts for 50% of R14's evidence
- **Dependance on red-teaming/external experts**
  - Gray Swan, FAR.AI, CAISI, UK AISI and benchmarks without disclosure



Conclusion

Threats to Validity & Future Work

- Current system cards fail to properly specify AI safety requirements, exhibiting vagueness, uneven evidentiary support, and poor traceability
- → To real-life harm and consequences
- Modelling these documentations can expose these gaps

Threats to Validity

- Internal: Possible claim-extraction bias (mitigated via cross-review with a third expert)
- External: Specific to GPT-5 system card, but consistent with preliminary work on GPT-OSS, Claude 3, and Gemma 3N

Future Work

- Assessing regulatory compliance
- Interviewing system-card authors to validate requirement interpretations

Thank You!

My Website

JPIPE
EST. 2023

Reproduction package