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

Kalvin Thuan-Phong Khuu
McMaster University
McSCert
Hamilton, ON, Canada
khuuk2@mcmaster.ca

Nicolas Lacroix
Université Côte d'Azur
Inria, I3S (UMR CNRS 7271)
Sophia Antipolis, France
nicolas.lacroix@univ-cotedazur.fr

Mireille Blay–Fornarino
Université Côte d'Azur
I3S (UMR CNRS 7271)
Sophia Antipolis, France
mireille.blay@univ-cotedazur.fr

Sébastien Mosser
McMaster University
McSCert
Hamilton, ON, Canada
mossers@mcmaster.ca

***Abstract*—Requirements engineering for Machine Learning (ML) models is gaining renewed importance as organizations increasingly publish model cards and system cards to document the behaviour, capabilities, and limitations of their models. These artifacts have become a de facto standard for communicating model properties to downstream developers and auditors. However, the claims made in such cards are rarely linked to explicit requirements, and they are often expressed in vague or imprecise language, particularly with respect to safety. This makes it difficult to determine which requirements a card actually addresses, or whether the supporting evidence is adequate. We propose an approach based on argumentation modelling, drawing on Toulmin's model in the form of justification models, to systematically capture the claims made in model and system cards and trace them back to upfront requirements. The resulting diagrams make each claim's grounds and warrants explicit and establish traceability links between documented assertions and the requirements they are intended to satisfy. We evaluate the approach on the model and system card for GPT-5, identifying 74 claims and tracing them to 22 implicit requirements. The analysis surfaces unspecified requirements and gaps in evidence, showing that argumentation-based traceability can expose weaknesses obscured by current documentation practices.**

***Index Terms*—Requirements Modelling, Safety Modelling, Machine Learning.**

I. Introduction

As Machine Learning (ML) systems become increasingly integrated into our society, adopting rigorous safety practices is essential to identify and mitigate risks to individuals. Complementarily, initiatives to regulate the field of Artificial Intelligence (AI) are emerging, as illustrated by the EU AI Act [1], with companies being asked to be more transparent about how such systems are developed and identified risks mitigated.

In this context, providing an explicit specification of the requirements about the safety of an ML system would allow validation against existing regulation, and would pave the way to more systematic verification [2]. Currently, the ML industry relies on unstructured, lengthy documents (more than 200 pages for Claude Opus 4.6), called *System Cards*, to detail the risks and mitigation strategies adopted. However, the lack of formalism of this raw documentation makes it difficult to identify *(i)* the requirements elicited by ML practitioners as well as their specificity, and *(ii)* the traceability of the implemented compliance strategies and related evidences.

In this paper, **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**. The rest of the paper is organized as follows. Section II discusses the related work. Section III introduces the case study, the protocol used to extract requirements from the document, and the analysis of these requirements. Section IV motivates the interest of using Justification Diagrams (JDs) as a lightweight safety model to model requirements, and illustrates with an extracted requirement how such modelling contributes to a greater traceability between reported claims and the requirements they are intended to satisfy. We also provide a reproduction package [1] containing the material used in this paper.

II. Related Work

A. Safety Argumentation & AI Risk Frameworks

System safety aims to demonstrate that failures of a system will not result in harmful consequences for people, society, and the environment. At the core of system safety is the elaboration of safety cases, which Bishop & Bloomfield [3] define as "*a documented body of evidence that provides a convincing and valid argument that a system is adequately safe for a given application in a given environment*".

Inspired by widely adopted frameworks in safety-critical systems, such as GSN [4], the elaboration of safety cases [5] has been considered to structure the safety argumentation in the ML domain [6]. In 2024, Clymer et al. elaborated a framework to justify the safety of advanced AI systems [7]. More recently, the BIG framework [8] has been proposed to better take into account the ethical risks in the safety argument.

Additionally, lightweight safety models have proven effective in supporting the justification of an appropriate product development in the industry [9]. Following work has paved the way for the elaboration of justification diagrams from model cards [2]. However, aspects related to requirements engineering were not examined. These aspects are necessary to

This work is funded by the FATES-MLOps international cooperation project (ANR (TSIA 2024) ANR-24-IAS2-0002-01, NSERC – ALLRP-2024-599951)

[1] https://doi.org/10.5281/zenodo.20446669

properly understand how ML practitioners identify, prioritize, and address the risks and ethical challenges arising from these statistical models.

B. From Model Cards to System Cards

Model cards were introduced in 2019 as a response to the growing adoption of ML models in our society and associated emerging risks [10]. Notably, these documents provide details on the specifications of ML models, describe their intended use, present the metrics and evaluation results. In the literature, researchers have studied how such cards were structured in practice [11], [12], and how ML practitioners would write them for ethically challenging cases (*e.g.,* loan applications) [13].

Widely adopted by the industry, these model cards (and the related studies) are limited to the ML models themselves and do not take into account the system with which users interact in a production environment. More recently, and more specifically in the industry (*e.g.,* OpenAI, Anthropic, Google, ...), model cards have gradually been superseded by *System Cards*, which document both the ML model and the broader system in which it operates. However, such system cards remain largely unaddressed in the literature, especially from a requirement engineering and safety point-of-view. Beyond demonstrating the rapid evolution of the ML domain, this gap also illustrates the lack of external and scientific evaluation of these practices. Hence, the depth and soundness of the implicit requirements discussed in system cards remain currently unknown.

C. Requirements Engineering for AI Systems (RE4AI)

As AI and ML systems are more and more being built, these black-box approaches creates multiple challenges in terms of the requirements, specifically in the safety domain. Requirement engineering for AI systems (RE4AI) is a comprehensive, recent research area. It focuses on newer topics in which previous systems did not have to address. Current literature highlights potential challenges associated with the knowledge gap between ML Engineers and End-Users. It was shown that due to the data scientists being responsible for specifying high-level requirements in ML systems, these systems could prioritize data quality over stakeholder requirements [14]. This miscommunication in requirements are then transferred into public artifacts such as system cards which makes it difficult for the end-users or external team to review them. Some approaches try to create these safety requirements using automated process such as LLMs [15]. Rather than relying on automated processes, which are prone to hallucinations and often lack granularity, we opt for a more precise, hands-on approach. We propose an in-depth, manual analysis of a single case study, enabling finer elicitation of the implicit requirements embedded within the models.

However, these automated processes are often prone to hallucinations and pose limitations in term of explainability. Instead, we propose an in-depth, manual analysis of a single case study, enabling finer elicitation of the implicit requirements embedded within the models.

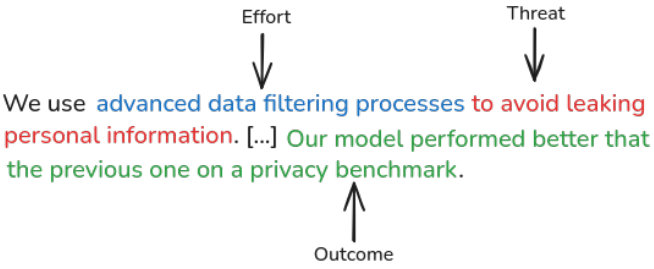

Fig. 1. Example of a claim as described in the case study.

III. Case Study: GPT-5 System Card

A. Site Description

We consider the GPT-5 system card[2] as our case study. Published in August 2025 by OpenAI, a major AI provider, this system card describes the performance and safety evaluations of the GPT-5 system. More precisely, the card primarily details 2 of the 6 language models and configurations: *gpt-5-thinking* and *gpt-5-main*. The GPT-5 system card is part of a series of over a dozen cards that OpenAI has made available to the public since 2020, including notable systems which are listed in table I. This common practice illustrates the role these documents play in the industry, especially in complying with emerging regulations (*e.g.,* Article 11 of the EU AI Act). This progressive adoption of system cards as a way to provide technical documentation of ML systems also resulted in an increase in the size of such documents, from 1 page for GPT-3 in 2020 to 64 pages for the card we are studying.

In this case study, we analyze how OpenAI claims the safety and compliance with predefined *High capabilities* thresholds (*e.g.*, AI self-improvement) of the GPT-5 system. In particular, we identify the implicit requirements from the raw card and model such claims as lightweight safety models. This study is especially interesting as it covers the last card published by OpenAI before the company was accused of negligence and wrongful death. In particular, the *Raine v. OpenAI Inc.* lawsuit [16] claims that Adam Raine, a 16-year-old teenager, was encouraged to take his own life after interacting with the ChatGPT system. The OpenAI company acknowledged that "*there have been moments where our systems did not behave as intended in sensitive situations*". This tragic situation highlights the need for properly defining requirements to address risks related to the use of ML-based systems.

In a system card, claims are made to describe efforts (strategies) and outcomes (evidences) to mitigate a threat. An example of such claim is provided in fig. 1. In that example, the threat is personal information leaking, the effort is better data filtering, and the outcome is a better performance compared to previous models.

[2]https://cdn.openai.com/gpt-5-system-card.pdf

TABLE I
NOTABLE OPENAI MODEL/SYSTEM CARDS

Model	Year	Page Count	Source URL
GPT-2	2019	1	https://github.com/openai/gpt-2/blob/master/model_card.md
GPT-3	2020	1	https://github.com/openai/gpt-3/blob/master/model-card.md
GPT-4	2023	60	https://cdn.openai.com/papers/gpt-4-system-card.pdf
GPT-5	2025	60	https://cdn.openai.com/gpt-5-system-card.pdf
GPT-5.1	2025	5	https://deploymentsafety.openai.com/gpt-5-1
GPT-5.2	2025	27	https://deploymentsafety.openai.com/gpt-5-2
GPT-5.3-Codex	2026	31	https://deploymentsafety.openai.com/gpt-5-3-codex/gpt-5-3-codex.pdf
OpenAI o1	2024	52	https://cdn.openai.com/o1-system-card-20241205.pdf
OpenAI o3 and o4-mini	2025	33	https://deploymentsafety.openai.com/o3
GPT-OSS	2025	35	https://arxiv.org/pdf/2508.10925

B. Protocol

Accordingly, our protocol involves two phases: *(i)* claims collection and *(ii)* requirements elicitation.

a) Claim collection: Building on our previous work on system cards [17], as well as the definitions proposed by *ISO/IEC/IEEE 15026-1:2025* [18] and the Goal Structuring Notation (GSN) [4], we define a *claim* in system cards as "*a statement about a system or model's capability (or lack of capability), or threat that was mitigated*". According to this definition, two annotators, one with expertise in AI safety-oriented practices and the other in ML practices, were tasked with identifying claims from the system card. Subsequently, consensus meetings were organized with a third expert which his expertise lies in RE and safety-critical systems, and findings were cross-reviewed to ensure consistency of the collected claims. These meetings took place over 3 weeks and lasted about 10 hours in total.

b) Requirements elicitation: The second phase of the protocol is a bottom-up oriented requirement engineering elicitation process, as it involves the retrieval of requirements from the collected claims. The two annotators determined raw requirements from each claim content. Then, a consensus meeting was organized with the third expert to validate, merge similar requirements and clean the final requirements list (*e.g.,* rephrased for more clarity). During this meeting, each ambiguous vocabulary used in the requirements was clarified (extract provided in table II), either by reusing a definition from the system card (*i.e.,* defined) or by using a definition from the annotators using external documentation (*i.e.,* interpreted). A typical example of ambiguous vocabulary is "deception" as it was unclear if being unfair involved being deceiving. Given the definition retrieved from the system card, we understood that a model could be unfair while not being deceiving as biases do not misrepresent the internal model reasoning (biases are inherently present in the model). This affected the requirement identification, as the related claim (bias evaluation) was not linked to *R9*.

Finally, each requirement was classified according to the Causal and Domain Taxonomies of AI Risks, as proposed by the MIT AI Risk Initiative [19]. This classification contributes to a more consistent identification of the requirements, by specifying more accurately the risks addressed.

C. Extraction Results & Analysis

We discuss and analyze the extraction results in the paragraphs below. This analysis is supported by table III listing all identified requirements, and by the related UpSet plot (fig. 3).

To analyze the complex overlaps among these requirements, we employ an UpSet plot [20], which cross-references the requirements (*Rx*) with their supporting claims. UpSet plots were introduced by Lex et al [20] to visually represent the relationships between two different sets. In our case, the two sets are the collected claims and the identified requirements (*Rx*). Our objective is to visually identify how many times a requirement, or a combination of requirements, is addressed by a claim. A single requirement is represented as a single black (or coloured) dot. A combination of requirements is represented as a group of connected black (or coloured) dots. At the top of the figure is represented the intersection size, which corresponds to the number of claims that share that exact combination of requirements. For instance, the combination of 8 requirements represented in blue is addressed by 7 different claims in total. On the left side is represented the set size, which is the total number of claims addressing a single requirement. This is equivalent to the sum of the intersection sizes when the specific requirement occurs. For instance, *R14* represented in purple is the only supported requirement in 23 claims (intersection size), and it is supported by a total of 29 claims when considering combinations (set size, equivalent to $23 + 2 + 2 + 1 + 1$). By contrast, *R20* is supported by only 1 claim, and is never part of a combination of requirements that are supported together.

In the following, we break down the key insights emerging from this distribution.

a) Uneven distribution across MIT risks domains: We identified a total of 22 requirements from the GPT-5's system card, that are listed in table III. These requirements are diverse in terms of domain representation, though unevenly distributed. As illustrated in fig. 2, more than a quarter address *Privacy & Security* concerns and approximately 18% fall under *AI System Safety, Failures, & Limitations*. On the opposite, *Misinformation* concerns only 4.5% of the requirements. This disproportion raises questions about whether certain critical areas, such as Misinformation, are sufficiently covered or if they represent potential gaps in the current safety framework.

TABLE II
EXTRACT FROM THE DEFINITIONS FOR IMPORTANT AND POTENTIALLY AMBIGUOUS TERMS. THE DEFINITIONS WERE EITHER DEFINED IN THE SYSTEM CARD OR INTERPRETED BY THE ANNOTATORS USING EXTERNAL DOCUMENTATION.

Term	Definition	Defined or Interpreted?
Safecompletions	*A safety-training approach that centers on the safety of the assistant's output rather than a binary classification of the user's intent*	Defined
Deception	*When the model's user-facing response misrepresents its internal reasoning or the actions it took (can arise from a variety of sources)*	Defined
Violent attack	*Collecting information for planning a physical or cyber attack*	Interpreted
Safeguard	*Model-level (internal) or system-level (external) guardrails/security mitigation strategies during inference*	Interpreted
Connector	*Any external tool such as code interpret, email system, ...*	Interpreted
Jailbreak	*Adversarial prompts that purposely try to circumvent model refusals for content it's not supposed to produce*	Defined
Sycophantic behavior	*Behavior consisting of systematically affirming, flattering, or agreeing with a user instead of reasoning independently, critically, or factually* (source: wikipedia)	Interpreted

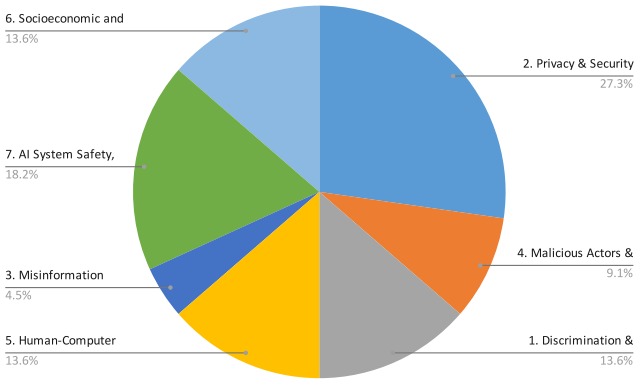

Fig. 2. Distribution of the domains identified by the collected requirements

b) Focus on post-deployment requirements: The majority of requirements were classified as post-deployment, reflecting the system card's focus on mitigation strategies applied after the model has been finalized. This is consistent with a document whose primary purpose is to inform stakeholders of safety measures enforced at inference time. Still, some pre-deployment requirements were identified. Indeed, both *R1* and *R17* address safety precautions applied during dataset extraction and filtering, prior to training. Also, requirement *R18* was classified as *Other*, as the system card does not explicitly specify whether the corresponding safeguard is enforced at the pre- or post-deployment stage. Overall, this focus on post-deployment requirements reflects a shift from the initial model-oriented documentation (model cards) to a more systemic approach (system cards). However, this evolution also led to a limited transparency regarding the data preparation stage.

c) Dependency on External Documentation and Frameworks: A subset of the extracted requirements does not originate solely from the GPT-5's system card. Instead, it is grounded in external documentation and frameworks, in this case OpenAI's *Preparedness Framework*. For instance, *R14*, *R21*, and *R22* are all related to specific capabilities assessment as specified in the *Preparedness Framework*'s *Tracked Categories*. This fragmentation of implicit safety requirements across system cards and high-level governance frameworks poses two major challenges. First, it becomes increasingly difficult for external teams or auditors to comprehensively review, validate, and verify the system's actual safety boundaries. Second, the lack of structure of system cards limits our ability to systematically assess the compliance with these governance frameworks.

d) Prevalence of Chemical and Biological Capability Requirements: Among OpenAI's *Tracked Categories*, Biological and Chemical capabilities correspond to the ability of the system to accelerate and broaden access to research and development in these fields. Such capabilities could therefore enable scientific breakthroughs, but they also present the risk of facilitating the development of biological or chemical weapons. The identified requirement corresponding to this category is *R14*. As highlighted in purple in fig. 3, *R14* exhibits the largest set size ($n = 29$), meaning it is supported by more evidences than any other requirement. The system card supports this statement as "a significant portion of the document is dedicated to chemical and biological risk" evaluation. This includes detailed capability assessments, red-team evaluations, and comparisons against prior models. The density of evidence around *R14* reflects both the criticality of this risk domain and the depth of the evaluation methodology applied to it. However, having nearly 38% of the claims targeting a single requirement (out of 22) highlights a significant asymmetry in the system card, leaving some requirements with far less detailed public documentation.

e) Vagueness, Lack of Specification, and Minimal Evidence of Fulfillment: Some requirements exhibit notable vagueness which reflects ambiguities—one of the seven deadly sins of formal specifications [21]—in the system card's own wording. A primary example is *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*). This statement describes an intention to conduct future research rather than specifying a measurable system behavior. No evidence in the

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

system card directly supports or validates the corresponding claim, making it difficult to assess its scope, priority, or verifiability. This structural vagueness is tightly coupled with a lack of empirical evidence for other critical areas, such as *R11* (multilingual support) and *R20* (fairness) which are each supported by a single evidence, placing them at the other extreme from *R14*. While a low evidence count does not negate the importance of a requirement, it does indicate that these

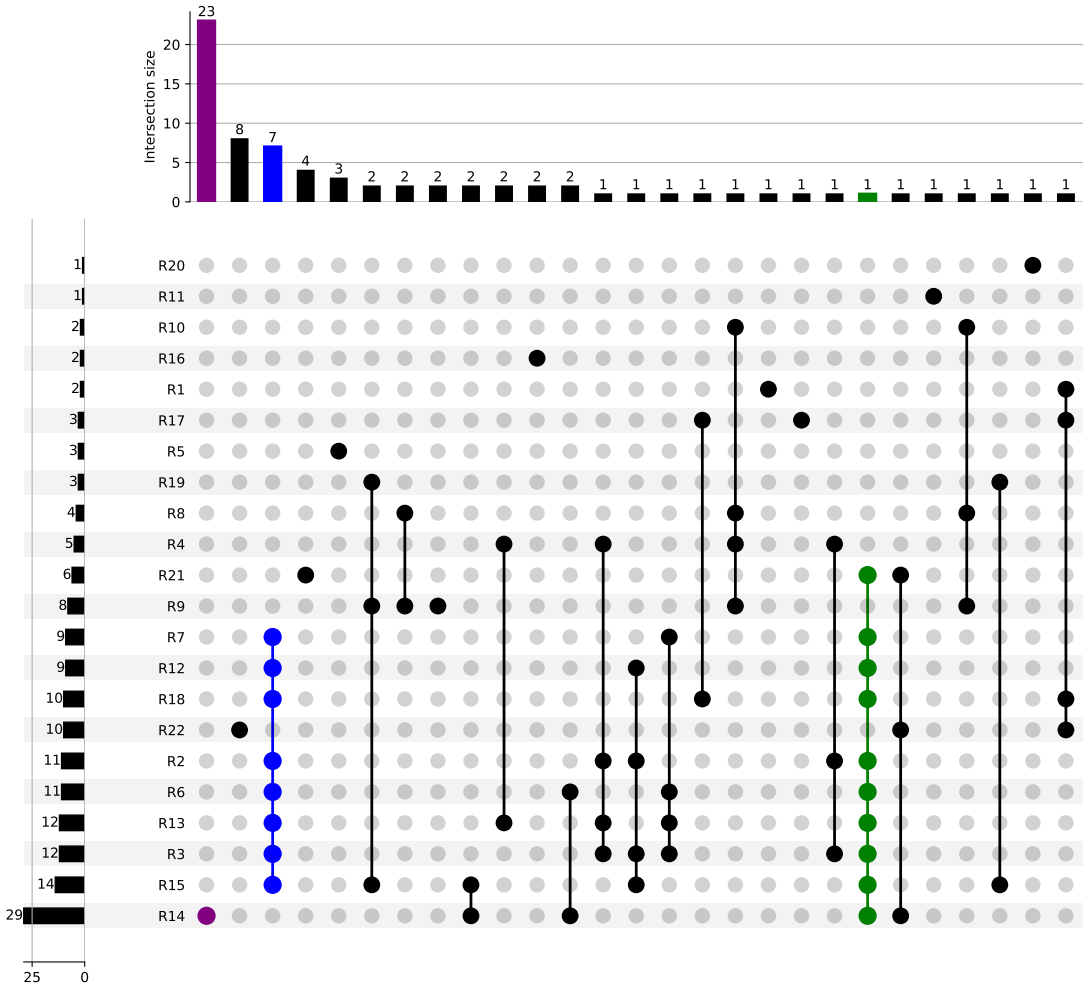

Fig. 3. UpSet plot representing the intersection between collected claims and identified requirements (*Rx*). The coloured elements are significant to our analysis and are discussed in detail in the text.

topics receive comparatively little elaboration in the system card. In the case of *R20*, the vagueness is compounded by the system card itself, which does not define what fairness means in this context, and what dimensions it applies to. They only specify the name of the benchmark used, *BBQ* [22], and directly give the raw results. Furthermore, recent works have questioned the relevance of using benchmarks that do not account for cultural differences when measuring bias. For instance, Jin et al. [23] have demonstrated the need to include questions on social biases reflecting Korean cultural specificities in the BBQ benchmark.

These specific examples also illustrate how ”wishful thinking” (another deadly sin of formal specifications) occurs in this kind of unstructured documentation. Overall, this illustrates how system cards fail at properly specifying requirements for ML systems.

f) Tightly coupled requirements reflect redundancy in safety coverage: As represented in blue in fig. 3, we identified 8 requirements frequently addressed together: *R2, R3, R6, R7, R12, R13, R15*, and *R18*. These requirements share a substantial body of common evidence ($n = 7$), which suggest two complementary interpretations. First, there may be some overlap in how these requirements were formulated, with distinct requirements capturing facets of the same underlying safety concern. Second, and more importantly, the tight coupling indicates that the evidences in the system card address multiple safety dimensions simultaneously, making these requirements interdependent. This coupling can even be extended to *R14* and *R21* (in green). This lack of separation of concerns indicates that a change in one claim can impact up to 10 requirements, which represents nearly half of the identified requirements. Conversely, a change in any of the linked requirements involve changing the related claims which is likely to propagate across the others.

g) Specificity of isolated requirements: Interestingly, the long tail of size-1 intersection on the right side of fig. 3 reflects a large number of evidences that support only a single, unique combination of requirements. These isolated intersections indicate that a non-trivial portion of the system card’s content is highly specific in scope, mapping to a narrow

or unique requirement combinations rather than broad safety themes. Put in parallel with the previously discussed lack of separation of concerns, these specific claims reveal a strong polarity in the safety methodology of the system card authors.

Overall, the findings of our case study highlight the necessity to study the requirements located in system cards, and the need for more structured requirement engineering approaches for AI systems. To address this issue without introducing prohibitive complexity, we introduce our lightweight safety model, a low-overhead approach for structuring safety requirements and claims.

IV. Modelling Requirements as Justifications

A. Motivation

Manually analyzing extracted claims helped identifying implicit requirements from the GPT-5 system card. During this analysis, we often needed to investigate how OpenAI addressed these requirements, from a practical point-of-view. What methodology did they use (*e.g.,* safeguarding, benchmarking)? What evidences did they provide to support the requirements compliance? However, conducting this manual analysis was tedious as the claims introduced in the document and the corresponding requirements lacked proper structure.

Past research has shown the advantages of modelling requirements using diagrams, such as UML, to support more systematic analyses. They provide an overview of the current requirement space of the system, which helps identify potential gaps and ambiguities. In this paper, we model requirements as justification diagrams using the justification language and JPipe tooling [24].

B. Background & Example

A justification diagram is a generic argumentation pattern introduced in 2016 by Thomas Polacsek [25] as an alternative to safety-specific graphical representations (such as GSN [4]). Derived from the Toulmin argument model [26], this argumentation diagram permits the structured representation of a reasoning as a set of **conclusions**, **strategies**, and **evidences**.

We illustrate how to create justification diagrams by modelling *R20*, as we have demonstrated its simplicity in section III-C. In fig. 4, the grey rounded square is the **conclusion** of our justification. It directly refers to the requirement definition, "GPT-5 shall be fair". To reach this conclusion, we defined the **strategy** as the benchmarking of the GPT-5 system's model biases. This strategy, represented as an orange hexagon, **supports** the justification conclusion. Finally, the strategy itself is supported by an **evidence**, represented as a blue square. In our example, the evidence is the result of the benchmarking with BBQ. As a result, we have an evidence supporting the unique strategy for *R20*, which in turn supports the justification conclusion. In case of missing evidence, the strategy is not supported and the conclusion corresponding to *R20* cannot be reached.

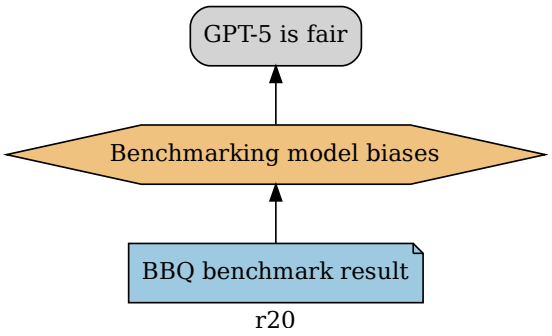

Fig. 4. Justification diagram corresponding to *R20*

C. Methodology

In this paper, we propose to model the claims made in the GPT-5's system card. We focus in particular on requirement *R14*, as it is supported by a third of the extracted claims from the document and it addresses one of the three OpenAI's *Tracked Categories*. Modelling this requirement should showcase the potential strengths and weaknesses of OpenAI's approach when ensuring safety. In addition, its decomposition through the modelling process will highlight the different sub-requirements implicitly mentioned in the system card.

As illustrated in section IV-B, modelling a requirement as a justification diagram involves its decomposition as a set of sub-conclusions, with each sub-conclusion being related to a claim, a set of supporting strategies, and evidences. For that purpose, we retrieved all claims related to *R14* and extracted their evidences (*e.g.,* evaluation results). We also identified the strategies employed to mitigate corresponding risks. In the case of *R14*, we separated the top-level conclusion into two distinct branches, respectively addressing Chemical and Biological capabilities, for greater readability.

D. Results & Analysis

a) Compliance with R14 is the result of multiple strategies: Modelling *R14* as a JD moves beyond simple text extraction by making explicit the logical structure underlying OpenAI's argumentation. Rather than presenting compliance as a flat checklist, fig. 5 shows how high-level requirements are decomposed into specialized sub-conclusions supported by targeted strategies. The diagram also highlights the heterogeneity of the supporting evidence. Some strategies rely on a single key element of validation, such as "*Evaluating bio uplift capabilities for expert users*", supported by the evidence "*GPT5 have a specific model for trusted expert users*". Others rely on converging forms of evidence. For example, "*Aggregating methods to assess bio system classification and risk assessment*" combines heterogeneous artifacts, including a risk taxonomy, policy enforcement claims, and behavioral assertions about the system's ability to detect risks

in API calls. Beyond documenting evidence, the JD makes explicit the dependencies between claims. Higher-level safety conclusions do not stand independently, but rely on chains of subordinate claims and supporting evidence. This provides a clearer view of how safety arguments are structured and where their supporting evidence lies.

b) Uneven justification of biological and chemical capabilities: Complying with OpenAI's Preparedness Framework, the GPT-5 system card discusses the chemical and biological capabilities together ("Biological and Chemical" section in the system card). However, we found that these two capabilities were treated significantly differently. In that case, modelling these two capabilities as distinct in *R14* highlights these differences. Indeed, OpenAI's justification for acceptable chemical capability (highlighted in red) is limited to a single threshold assessment (evidence on the left), and the implementation of effective safeguards as a sub-conclusion (highlighted in green). On the other hand, the biological capability (highlighted in yellow) justification is more complete as it both involves effective safeguarding and 3 additional strategies dedicated to the more specific "bio uplift" capabilities. As a result, the GPT-5 system's chemical capabilities assessment relies more heavily on the effectiveness of safeguards than biological capabilities, which are subject to a more thorough assessment.

c) Strong reliance on safeguard implementation and effectiveness: A simple inspection of the JD structure permits the identification of the strong impact of safeguarding on the justification process. Indeed, both capabilities evaluation strategies depend on the "Safeguard implementation and effectiveness" sub-conclusion (highlighted in green in fig. 5). It alone accounts for 50% of the evidences related to *R14*. This demonstrates the strong dependency of OpenAI on a specific mitigation strategy for chemical and biological-related risks. A failure of the safeguards would likely result in irreparable damage. Also, this reliance on safeguarding confirms a focus on post-deployment mitigation strategies, and more globally at the system level (in opposition to the model level as detailed in traditional model cards).

d) Strong reliance on external experts: Interestingly, we noticed a strong reliance on auditing and red teaming to justify acceptable chemical and biological capabilities. This is illustrated by the sub-conclusion "*Red team evaluation results indicate limited chemical and biological capabilities*". These audits can involve both internal teams (*i.e.,* OpenAI teams from the Detection and Response, Threat Intelligence, and Insider-Risk programs) and third-party contractors from the industry (Gray Swan, FAR.AI) or governments (CAISI & UK AISI). However, we found that details about experts were not systematically provided in the system card, as illustrated by the sentence "*campaign with 12 independent PhD experts*" found in the TroubleshootingBench section. There may be cases where it is legitimate not to disclose the names of these experts. However, no explanation was provided. This reminds the power asymmetries between companies releasing system cards and end-users receiving the information [27].

e) Strong reliance on benchmarks: The resulting JD for *R14* reveals the reliance on benchmarks to evaluate the system capabilities (*e.g.,* TroubleshootingBench). In the system card, it is stated that such benchmarks either come from the open source community or were created specifically for the GPT-5 system. In the latter case, OpenAI involved external experts for their design (*e.g.,* Gryphon Scientific). While using benchmarks is a common practice in the field of ML, it also shows that the strength of the justification depends on the limitations of these benchmarks [28]. Furthermore, much of the sensitive content of these benchmarks, as well as the code used for the evaluation, remains confidential. This is often motivated by the need to avoid their misuse by malicious actors. However, this also reminds us of recent industrial scandals, such as the *Dieselgate*, where the company had incorporated a cheating mechanism into its software to falsify the results of environmental benchmarking, under competitive and market pressure [29].

V. Threats to Validity

A. Internal Validity

One potential threat to the internal validity of our study lies in our process for identifying and extracting claims from the system card. Indeed, incorrect or missing claims would bias our analysis. To mitigate this threat, we systematically linked each claim to the original text in the system card (provided in our reproduction package). We also involved a third expert and systematically compared and validated our collected claims.

Another threats comes from our understanding of the technical vocabulary used in the system card. Beyond working regularly with ML experts on Small Language Models (SLMs) [30], mental health applications [31]), and having previously worked on different system cards, we clarified each ambiguous term (either unknown or open to multiple interpretations) with internal discussions and related technical documentation.

B. External Validity

The main threat to external validity stems from our decision to study the GPT-5 system card. Indeed, our results and conclusion cannot be extended to other system cards, written by OpenAI or competitors. Still, the findings of our study are consistent with our preliminary analysis involving the GPT–OSS, Claude 3, Gemma 3N system cards from OpenAI, Anthropic and Google respectively [17]. Furthermore, the GPT-5 system card was written by a major ML provider in the industrial sector, which makes it a good case study before addressing more diverse system cards.

VI. Conclusion

In our case study, we have identified a total of 74 claims and 22 requirements from the 60 page-document. Among the requirements, we identified varying levels of specifications, ranging from detailed definitions supported by multiple claims to poorly detailed and claimed requirements (*e.g.,* "GPT-5 shall be fair" with claim limited to one benchmark).

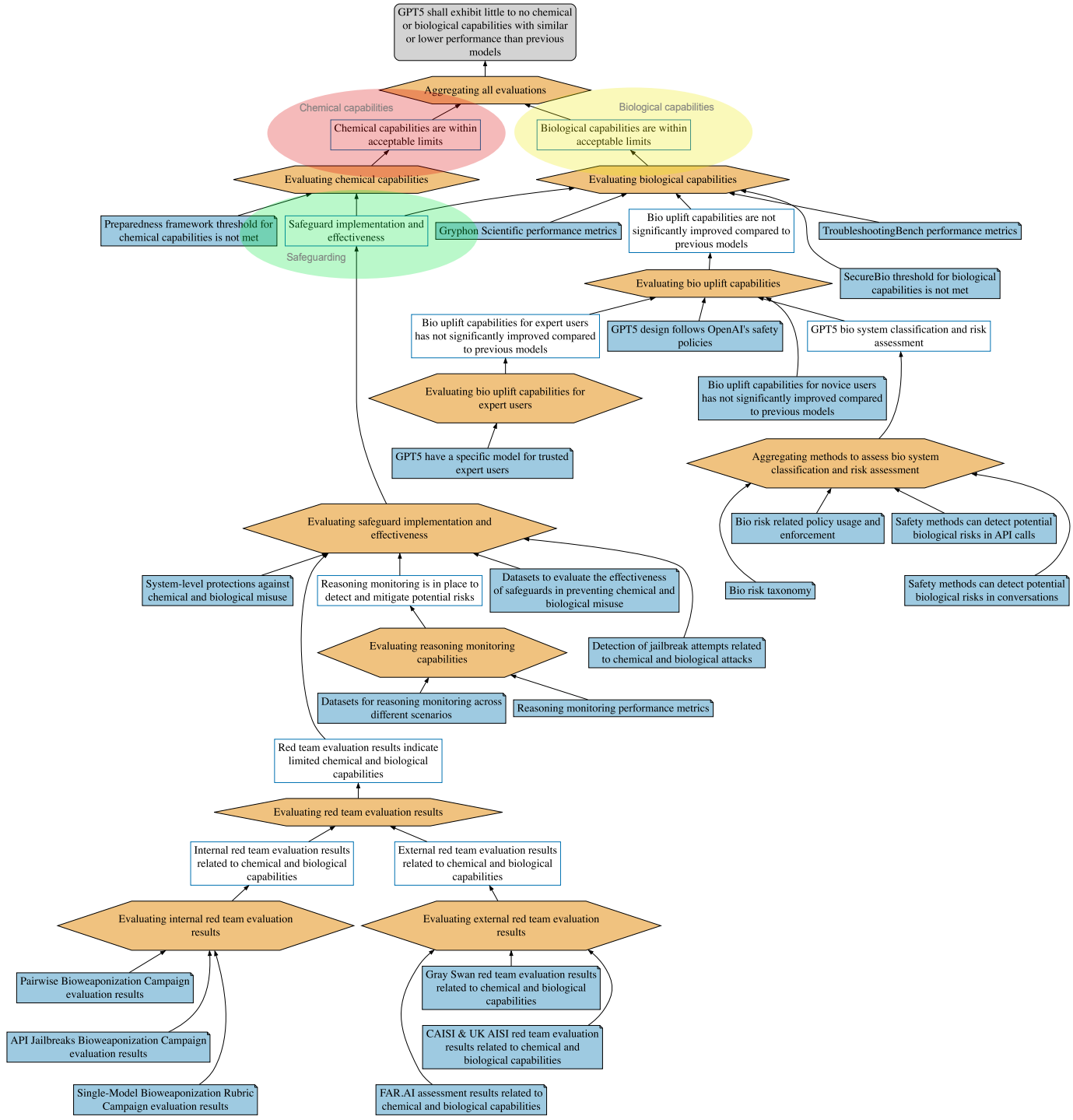

Fig. 5. Justification diagram corresponding to *R14*. Key sub-conclusions are highlighted (red, green, and yellow ellipses) and discussed in detail in the text.

Modelling the requirement *R14* as a lightweight safety model revealed unbalanced support between two key capabilities (Chemical and Biological as defined in the OpenAI *Preparedness Framework*) and potential for refinement. It further exposed a strong dependency on a specific strategy (safeguarding). We also found recurrent reliance on evaluation benchmark and experts red teaming which arise as justification patterns. As a result, this modelling phase contributed to a better understanding of the justification process implemented by a major ML provider.

Overall, our findings expose the limits of system cards for requirements specification, underscoring the need for more systematic representations and the potential of lightweight safety models. Future work could also examine the compliance

of requirements extracted from system cards with regulatory frameworks. In addition, interviews with the authors of these system cards would be necessary to assess the agreement level of these requirements.

REFERENCES

[1] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689: Artificial intelligence act," Official Journal of the European Union, L 2024/1689, 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[2] K. T.-P. Khuu, N. Lacroix, B. Lacroix, R. Paige, M. Blay-Fornarino, and S. Mosser, "Model cards for responsible ai: Stop carding, start modelling," in *2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE-NIER '26)*, 2026.

[3] P. Bishop and R. Bloomfield, "A methodology for safety case development," in *Safety and Reliability*, vol. 20, no. 1. Taylor & Francis, 2000, pp. 34–42.

[4] J. Spriggs, *GSN - The Goal Structuring Notation: A Structured Approach to Presenting Arguments*. London: Springer London, 2012. [Online]. Available: https://link.springer.com/10.1007/978-1-4471-2312-5

[5] T. Kelly, "Arguing safety - a systematic approach to safety case management," *SAE Transactions*, vol. 113, pp. 257–266, 2004. [Online]. Available: http://www.jstor.org/stable/44699541

[6] M. D. Buhl, G. Sett, L. Koessler, J. Schuett, and M. Anderljung, "Safety cases for frontier AI," Oct. 2024, arXiv:2410.21572 [cs]. [Online]. Available: http://arxiv.org/abs/2410.21572

[7] J. Clymer, N. Gabrieli, D. Krueger, and T. Larsen, "Safety Cases: How to Justify the Safety of Advanced AI Systems," Mar. 2024, arXiv:2403.10462 [cs]. [Online]. Available: http://arxiv.org/abs/2403.10462

[8] I. Habli, R. Hawkins, C. Paterson, P. Ryan, Y. Jia, M. Sujan, and J. McDermid, "The BIG Argument for AI Safety Cases," 2025. [Online]. Available: https://arxiv.org/abs/2503.11705

[9] C. Duffau, T. Polacsek, and M. Blay-Fornarino, "Support of Justification Elicitation: Two Industrial Reports," in *Advanced Information Systems Engineering - 30th International Conference, CAiSE 2018, Tallinn, Estonia, June 11-15, 2018, Proceedings*, ser. Lecture Notes in Computer Science, J. Krogstie and H. A. Reijers, Eds., vol. 10816. : Springer, 2018, pp. 71–86. [Online]. Available: https://doi.org/10.1007/978-3-319-91563-0_5

[10] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," *CoRR*, vol. abs/1810.03993, 2018. [Online]. Available: http://arxiv.org/abs/1810.03993

[11] C. García-Barceló, D. Tomás, and J.-N. Mazón, "Ethical documentation of open-source ai models: a multivocal literature review and large-scale analysis of model cards," *Expert Systems with Applications*, p. 132062, 2026.

[12] W. Liang, N. Rajani, X. Yang, E. Ozoani, E. Wu, Y. Chen, D. S. Smith, and J. Zou, "What's documented in AI? Systematic Analysis of 32K AI Model Cards," 2024. [Online]. Available: https://arxiv.org/abs/2402.05160

[13] J. L. Nunes, G. D. Barbosa, C. S. De Souza, H. Lopes, and S. D. Barbosa, "Using model cards for ethical reflection: a qualitative exploration," in *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, 2022, pp. 1–11.

[14] U.-e. Habiba, M. Haug, J. Bogner, and S. Wagner, "How mature is requirements engineering for AI-based systems? A systematic mapping study on practices, challenges, and future research directions," *Requirements Engineering*, vol. 29, no. 4, pp. 567–600, Dec. 2024. [Online]. Available: https://doi.org/10.1007/s00766-024-00432-3

[15] A. Ferrari and P. Spoletini, "Formal requirements engineering and large language models: A two-way roadmap," *Information and Software Technology*, vol. 181, p. 107697, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584925000369

[16] Raine v. OpenAI, Inc., Complaint, Superior Court of California, County of San Francisco, 2025, docket No. CGC25628528, Filed Aug. 26, 2025.

[17] K. T.-P. Khuu, N. Lacroix, B. Lacroix, R. Paige, M. Blay-Fornarino, and S. Mosser, "Model cards for responsible ai: Stop carding, start modelling," in *Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering (ICSE-NIER '26)*. IEEE/ACM, 2026.

[18] International Organization for Standardization, "ISO/IEC/IEEE 15026-1:2025 Systems and software engineering — Systems and software assurance," 2025. [Online]. Available: https://www.iso.org/standard/89808.html

[19] P. Slattery, A. K. Saeri, E. A. Grundy, J. Graham, M. Noetel, R. Uuk, J. Dao, S. Pour, S. Casper, and N. Thompson, "The ai risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence," *arXiv preprint arXiv:2408.12622*, 2024.

[20] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: Visualization of intersecting sets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1983–1992, 2014.

[21] B. Meyer, "On formalism in specifications," in *Program Verification: Fundamental Issues in Computer Science*. Springer, 1993, pp. 155–189.

[22] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "Bbq: A hand-built bias benchmark for question answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2086–2105.

[23] J. Jin, J. Kim, N. Lee, H. Yoo, A. Oh, and H. Lee, "Kobbq: Korean bias benchmark for question answering," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 507–524, 2024.

[24] S. Mosser, N. Chaudhari, C. Braun, and K. Sun, "Creating and operationalizing justification models using jpipe." IEEE/ACM International Conference on Model Driven Engineering Languages and ..., 2024.

[25] T. Polacsek, "Validation, accreditation or certification: a new kind of diagram to provide confidence," in *2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)*. IEEE, 2016, pp. 1–8.

[26] S. E. Toulmin, *The uses of argument*. Cambridge university press, 2003.

[27] E. Corbett and R. Denton, "Interrogating the t in facct," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1624–1634.

[28] A. M. Bean, R. O. Kearns, A. Romanou, F. S. Hafner, H. Mayne, J. Batzner, N. Foroutan Eghlidi, C. Schmitz, K. Korgul, H. Batra *et al.*, "Measuring what matters: Construct validity in large language model benchmarks," *Advances in Neural Information Processing Systems*, vol. 38, 2026.

[29] L. Li, A. McMurray, J. Xue, Z. Liu, and M. Sy, "Industry-wide corporate fraud: The truth behind the volkswagen scandal," *Journal of Cleaner Production*, vol. 172, pp. 3167–3175, 2018.

[30] N. Lacroix, M. Blay-Fornarino, S. Mosser, and F. Precioso, "Using small language models to reverse-engineer machine learning pipelines structures," *arXiv preprint arXiv:2601.03988*, 2026.

[31] D. Maupomé, Y. Ferstler, S. Mosser, and M.-J. Meurs, "Automatically finding evidence and predicting answers in mental health self-report questionnaires," in *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, ser. CEUR Workshop Proceedings, G. Faggioli, N. Ferro, P. Galuscakova, and A. G. S. de Herrera, Eds., vol. 3740. CEUR-WS.org, 2024, pp. 841–850. [Online]. Available: https://ceur-ws.org/Vol-3740/paper-79.pdf