

# RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting

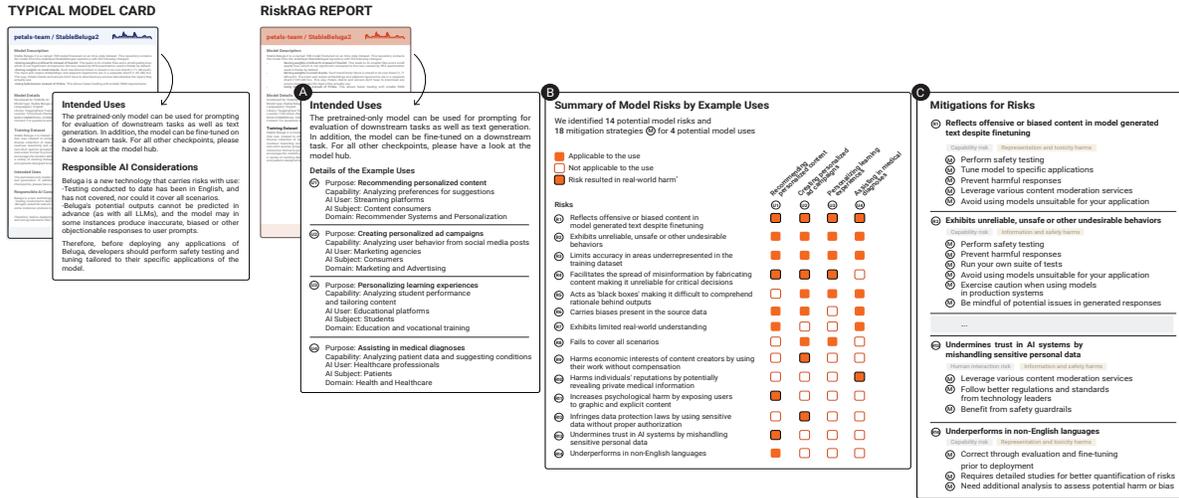
Pooja S. B. Rao  
 pooja.rao@unil.ch  
 University of Lausanne  
 Lausanne, Switzerland  
 International Institute of Information  
 Technology Bangalore  
 Bangalore, India

Sanja Šćepanović  
 sanja.scepanovic@nokia-bell-labs.com  
 Nokia Bell Labs  
 Cambridge, UK

Ke Zhou  
 ke.zhou@nokia-bell-labs.com  
 Nokia Bell Labs  
 Cambridge, UK  
 University of Nottingham  
 Nottingham, UK

Edyta Paulina Bogucka  
 edyta.bogucka@nokia-bell-labs.com  
 Nokia Bell Labs  
 Cambridge, UK  
 University of Cambridge  
 Cambridge, UK

Daniele Quercia  
 quercia@cantab.net  
 Nokia Bell Labs  
 Cambridge, UK  
 Politecnico di Torino  
 Turin, Italy



**Figure 1: Comparison of a typical model card (left), which omits risk discussions (86% do so), with a report generated using RiskRAG (right). The RiskRAG report includes exemplary model uses (A), a summary of risks by use case (B), and corresponding mitigations (C).**

## Abstract

Risk reporting is essential for documenting AI models, yet only 14% of model cards mention risks, out of which 96% copying content from a small set of cards, leading to a lack of actionable insights. Existing proposals for improving model cards do not resolve these issues. To address this, we introduce RiskRAG, a Retrieval Augmented Generation based risk reporting solution guided by five design requirements we identified from literature, and co-design with

16 developers: identifying diverse model-specific risks, clearly presenting and prioritizing them, contextualizing for real-world uses, and offering actionable mitigation strategies. Drawing from 450K model cards and 600 real-world incidents, RiskRAG pre-populates contextualized risk reports. A preliminary study with 50 developers showed that they preferred RiskRAG over standard model cards, as it better met all the design requirements. A final study with 38 developers, 40 designers, and 37 media professionals showed that RiskRAG improved their way of selecting the AI model for a specific application, encouraging a more careful and deliberative decision-making. The RiskRAG project page is accessible at: <https://social-dynamics.net/ai-risks/card>.

CHI '25, April 26-May 1, 2025, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Yokohama, Japan, <https://doi.org/10.1145/3706598.3713979>.

## CCS Concepts

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

## Keywords

AI risk, responsible AI, AI model, model cards, risk report, harm, incident

### ACM Reference Format:

Pooja S. B. Rao, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 26 pages. <https://doi.org/10.1145/3706598.3713979>

## 1 Introduction

As artificial intelligence (AI) becomes increasingly pervasive, identifying and reporting potential risks of an AI model is essential for its responsible and trustworthy development and use [18]. Ensuring AI safety is shared among stakeholders in the AI lifecycle [53]. By systematically reporting risks, AI developers can tackle ethical issues like fairness, accountability, and bias [32, 70], aligning AI models with societal values and reducing harm. Sharing risks and mitigation strategies enables application developers and organizations to enhance the safety and reliability of AI technologies. Lastly, transparent risk reporting helps consumers and technology users understand the potential risks of AI systems, enabling informed decisions [22], while also addressing broader societal considerations [71], and supporting compliance with regulatory frameworks such as the EU AI Act [14] and the U.S. AI Bill of Rights [30].

Model cards have become a de facto standard for reporting on AI models, widely adopted by both major tech companies and individual AI developers alike. For instance, HuggingFace<sup>1</sup>, one of the most popular model repository platforms, hosts 750K AI models, with 450,000 of them featuring model cards. Initially proposed by Mitchell et al. [48], model cards were introduced to standardize ethical reporting, clarify intended uses, and document the risks and limitations of AI models. Model cards encompass sections with technical information like model description, model usage, training and evaluation details, version and license, as well as the sections on intended uses, ethical considerations, risks and limitations, out-of-scope uses, and misuse.

Despite the intents of model cards, research reveals that only 17% of model cards briefly address issues related to bias or ethics [4, 37]. Our analysis of a more recent snapshot from July 2024 of 450K model cards finds this to be 14% (i.e., 64K model cards with risks reported). Moreover, a staggering 96% of these cards have their risk content copied from (i.e., identical to) an initial set of 2672 cards. Furthermore, most model documentation is insufficient for reasoning about the impact of model adoption, as the risk sections in these model cards are often criticized for being overly vague and generic, which restricts their practical application in decision-making processes [4, 15]. It has also been shown that anticipating the risks of an AI system or a model is a hard task even for practitioners and researchers with knowledge of AI [8, 19]. The increasing frequency

of real-world AI incidents and harms [45, 68] is likely partly due to the lack of transparency regarding risks associated with models deployed [4, 13]. Risks can be reported as “model-specific”, i.e., risks arising from the model’s unique capabilities or limitations (e.g., perpetuating harmful biases from training data), or “use-specific”, i.e., contextualized risks tied to specific applications (e.g., biases affecting attendees when transcribing virtual meeting).

Two key parallel areas of research attempt to address these challenges (Table 7). The first area focuses on enhancing *AI risk documentation* and developing supportive toolkits. Existing studies have investigated approaches to improve the usability and effectiveness of model cards such as introducing interactivity [15], adopting nudging formats like DocML [4], and treating AI documentation as a formal project deliverable [13] to encourage better documentation practices. These efforts have contributed to improved reporting of model-specific risks, particularly technical and capability-related risks. However, they often fall short of contextualizing these risks for specific uses [58]. This gap is becoming increasingly significant as legal frameworks, such as the EU AI Act, prioritize evaluating AI systems within their usage contexts [14]. To complement model documentation, researchers have introduced Risk Cards [17], a novel documentation format. While Risk Cards emphasize the importance of contextualizing risks for specific uses, they are designed to document individual risks separately. As a result, representing the full spectrum of risks for a single AI model would require multiple Risk Cards. Additionally, Risk Cards explicitly discourage documenting specific models, making them unsuitable for comprehensive model-level risk assessment.

The second area of research focuses on developing *tools for populating proposed documentation*. These tools are designed to help practitioners envision potential uses [29] as well as identify risks and harms [11, 69] associated with AI systems. However, existing tools do not consider the specific model underlying the AI system and therefore do not generate model-specific risks, i.e., those unique vulnerabilities or limitations tied to a particular model’s design, development, or deployment. For instance, while these tools might identify risks associated with text generation language models in general, they fail to distinguish risks unique to different models that could arise, for example, from the specific data they were trained on (e.g., not safe for work images). For model cards in particular, a retrieval-augmented generation (RAG) [36] solution called CardGen [39] has been proposed to assist in filling them in. However, CardGen adheres to existing textual formats when populating risk sections, reinforcing the same shortcomings that have been criticized [4, 37] rather than addressing them [4, 15].

To address the challenges of AI model risk reporting, we built upon and contributed to both areas of research. In doing so, we made three key contributions:

- (1) *Enhanced AI Model Risk Report (§3)*. Model reporting, including risk reporting, is important for a variety of stakeholders in the three AI phases of development, deployment, and use [48], as it facilitates effective communication among groups with diverse roles [27]. The AI developers who typically create model cards are our key target audience. We identified five key design requirements for AI model risk reporting through a literature review and an iterative co-design study

<sup>1</sup><https://huggingface.co/>

with 16 experienced AI developers. These requirements are: identifying diverse model-specific risks, clearly presenting risks, prioritizing them, contextualizing for real-world uses, and offering actionable mitigation strategies.

- (2) *Automatically Pre-Populating the Report with RiskRAG (§4).* We developed a solution to support developers in producing actionable and understandable model risk reports that align with the previously identified five design requirements. Our approach is data-driven, leveraging existing databases of human-written risks and limitations in AI models (i.e., those reported by developers in model cards, and those reported by media professionals in AI incident reports). Specifically, from an initial 450K model cards from HuggingFace, we compiled a dataset of 2672 cards containing unique risk sections and incorporated 649 real-world AI incidents from the AI Incident Database [45] to capture a diverse spectrum of risks. RiskRAG utilizes a RAG framework to retrieve relevant risks and mitigation strategies from these sources, presenting them in a clear and structured format (Figure 1 shows an example).
- (3) *Empirically Validating the RiskRAG Report (§5).* We first conducted a baseline evaluation (§5.1) to validate the quality and relevance of RiskRAG content against existing high-quality model cards written by developers. Next, we ran a preliminary user study (§5.2) with 50 AI developers tasked with advocating for the adoption of an AI model in a high-risk hiring application. Participants preferred using the RiskRAG report for this task compared to a baseline model card (74%), and rated it higher in meeting all the design requirements. Although AI developers typically create model cards, these reports are consumed by a broader audience, including non-technical users. Hence, in the final user study (§5.3), we also involved non-technical users to assess RiskRAG’s broader relevance and its lower bound performance for general applicability. This study involved 38 developers, 40 designers, and 37 media professionals, who were tasked with selecting between two AI models for media industry tasks. Across all groups, RiskRAG improved the argumentation in the model selection explanation, and encouraged more cautious decision-making. Participants consistently preferred the RiskRAG report, citing its clarity and support in decision-making.

Our solution improves risk reporting by providing: (1) an enhanced format, and (2) reducing the effort required for developers to document high-quality risks in this format. Importantly, we see RiskRAG *not* as a definitive solution but as *a tool to assist AI developers in creating effective risk reports*.

## 2 Related Work

Two parallel areas of related research on AI risk reporting (Table 7) focus on: (1) developing formats for documenting AI risks (§2.1), and (2) creating automated tools to assist in filling in such documentation (§2.2).

### 2.1 AI Risk Documentation

Various forms of documentation have been proposed to support responsible AI (RAI) practices: from model documentation (e.g., model cards [48]), dataset documentation (e.g., datasheets for datasets [24], data statements for NLP [3]), documentation for the purpose of using AI (i.e., ethics sheets for AI tasks [49]), to recent documentation for AI risks (i.e., Risk Cards [17]).

Mitchell et al. [48] introduced model cards for transparent model reporting, covering ethical considerations like sensitive data, potential risks, unintended uses, and mitigation strategies. Over time, model cards have become a standard, endorsed by regulations [14], governance frameworks [63], and major tech companies [27].

**Challenges.** However, model documentation format is still evolving, with some sections more frequently completed than others. An analysis of 32K model cards from HuggingFace revealed that only 17% of all cards and 39% of the top 100 most downloaded included sections on risks and limitations [37]. Another study reported similar findings when qualitatively analysing model cards also from GitHub and organizational websites [4]. Similarly, an analysis of dataset datasheets [24] found that while dataset descriptions are usually complete, considerations for appropriate data usage receive minimal attention [73]. Further analysis by Liang et al. [37] showed that risk sections of model cards typically address data and model limitations, focusing primarily on technical aspects. As a result, developers often find current risk sections ambiguous and lacking specificity [15], leading to a gap between what users need and what is provided [4].

**Solutions.** To tackle these issues, Crisan et al. [15] explored design choices for an interactive model cards version, while Bhat et al. [4] introduced DocML, a tool for improving documentation practices through nudging and traceability. Similarly, to incentivize risk reporting, Chang and Custis [13], suggested making model documentation a mandatory AI project deliverable. Additionally, Kennedy-Mayo and Gord [34] proposed restructuring the ethical considerations section to clearly outline regulatory, reputational, and operational risks. Beyond enhancing model cards, Derczynski et al. [17] introduced Risk Cards, a new type of RAI documentation specifically designed to address risks. Risk Cards are intended to complement other documentation by enabling cataloging of individual risks.

**Research gap.** To sum up, while previous research has explored improving AI model documentation, it has primarily focused on general practices rather than specifically addressing risk reporting. The efforts that focused on risk reporting such as [34] still fall short of contextualizing risks for specific uses. We argue that only when contextualizing them to uses do the other types of risks, such as human-interaction or systemic [68], begin to emerge. Moreover, legal frameworks such as the EU AI Act prioritize evaluating AI systems within their usage contexts [14]. Risk Cards do contextualize risks for specific uses; however, they can only serve as a complementary form of documentation rather than a substitute for model cards, as they focus on individual risks rather than models, and explicitly discourage documenting specific models in relation to those risks. To address these limitations, we derived key design requirements for an effective solution through a literature-informed co-design process.

## 2.2 Tools for Populating AI Risk Documentation

The need for reporting AI risks both at the level of models and specific uses is partly driven by standards like the NIST AI Risk Management Framework [51] and regulations like the EU AI Act [14], which mandate risk documentation based on the particular use and context [26, 31]. Consequently, various AI impact assessment reports [1, 6, 65] and cards [25] have been proposed to help AI developers prepare the required documentation, particularly for high-risk systems.

**Challenges.** Filling in this documentation demands envisioning the AI system’s uses and risks. Besides, AI risk assessment challenges [4, 13]. AI developers often struggle to envision specific uses and identify associated risks [29, 69].

**Solutions.** To address this, several semi-automatic tools have been proposed. Herdel et al. [29] introduced a large language model (LLM)-based tool, ExploreGen to support developers in envisioning potential uses and assessing the regulatory risk associated with each. Buçinca et al. [11] proposed AHA!, a tool combining LLMs and crowdsourcing that assists in anticipating potential harms and unintended consequences before developing or deploying an AI system. Wang et al. [69] introduced FarSight, another LLM-based tool designed specifically to support prompt developers working with LLMs. Bogucka et al. [7] compiled risks of various AI uses that have led to real-world harms, presenting them in a visualization appealing to the broader public. All of these tools leveraged LLMs to identify potential uses or risks for a given AI system. Lastly, Liu et al. [39] introduced CardGen, a RAG pipeline that helps fill missing sections in model cards, including the risk ones, using information sourced from respective papers and GitHub projects.

**Research gap.** ExploreGen [29] produces only model uses, while AHA! [11], and FarSight [69] produce use-specific but not model-specific risks. For instance, they do not differentiate risks between two image generation models, e.g., one trained on NSFW (not safe for work) images, and another on safe images. The former model warrants highlighting risks related to generating abusive, violent, or pornographic content if misused, whereas the latter may not pose such risks. As we will demonstrate, RiskRAG makes this distinction (Appendix E). Moreover, most existing tools rely solely on LLMs, which struggle with domain-specific or knowledge-intensive tasks due to hallucinations, and a lack of grounding in the specific domain knowledge [33, 76]. In contrast, RAG [36] combines retrieval with AI-generated responses, reducing hallucinations and enhancing task-specific accuracy without additional training [23, 55].

While CardGen [39] employs RAG, and is designed to fill missing sections of model cards on HuggingFace, including risk-related sections, it does so by replicating the existing format, and with it, its limitations identified in prior studies [4, 37]. Additionally, many models on HuggingFace lack associated research papers or repositories, limiting CardGen’s effectiveness in generating risk-related content for these models. Unlike CardGen, RiskRAG populates a finer-grained risk report for all model cards, even those lacking associated papers or external repositories meeting our five identified design requirements.

## 3 Deriving Design Requirements From Literature and an Iterative Co-design Process

We derived the design requirements for effective model risk reporting by reviewing the literature to identify initial requirements (§3.1). We then expanded on these through a co-design study with AI developers (§3.2).

### 3.1 Design Requirements From Literature

To gather requirements for reporting risks of the AI models, we began with a recent literature review (Figure 2, Step 1A). Our aim was to create a foundational scaffold for a model card risk section that could be enhanced in subsequent co-design iterations. We sought to uncover major flaws in risk reporting by analyzing a selection of well-scoped, high-quality papers rather than a large set of papers identified by an exhaustive review. Prior literature has demonstrated that this approach effectively generates initial design considerations for artifact creation [6, 16].

We conducted a keyword search within the ACM Digital Library (DL) as shown in Figure 3, choosing this resource due to its inclusion of SIGCHI publications and proceedings from AIES and FAccT conferences, where we anticipated finding relevant papers. Our search targeted articles published from 2019 onward, marking the publication year of the first model reporting proposal, model cards [48]. The initial search yielded 326 articles. After removing extended abstracts, magazine articles, and other short-form papers or reports, 263 articles remained.<sup>2</sup> To ensure the relevance and quality of these papers, we applied the following inclusion criteria: (1) relevance to AI model risk reporting, (2) focus on the documentation practices of AI models, and (3) presentation of tools for documenting AI models.

Based on these criteria, we screened papers by title and abstract. This eliminated the majority of the papers focused on evaluating the limitations of the model or domain-specific papers like healthcare and education, narrowing our selection to six papers. Given the rapid development of research on ethical, responsible, and trustworthy AI, we also conducted a similar search in Arxiv to capture the latest studies, which added two unpublished papers to our list, both fairly referenced by other research attesting to their quality. The final selection of eight papers [4, 13, 15, 21, 34, 37, 48, 52] provided insights into the diverse aspects of AI risk reporting in model documentation, allowing us to synthesize initial design requirements for the risk report.<sup>3</sup> We conducted a full-text qualitative analysis of these eight papers. A thematic analysis [9, 10] was employed, focusing on sections relevant to risk reporting to identify key design requirements for reporting AI model risks. The first two authors distributed the papers between themselves and reviewed them to gain familiarity with their content. Using a bottom-up approach, we then coded the different sections from each paper, refining them as the coding progressed. This procedure resulted in 20 codes, which were organized into a thematic map and grouped into 6 sub-themes.

<sup>2</sup>Supplementary Materials (Supp. Mat.) are available on OSF at <https://osf.io/chjgp/>. We provide the entire list of papers in [Supp. Mat. 1.1](#).

<sup>3</sup>Refer to §2 for a review of these papers. Appendix A lists these papers and their relevance to risk reporting in model cards.

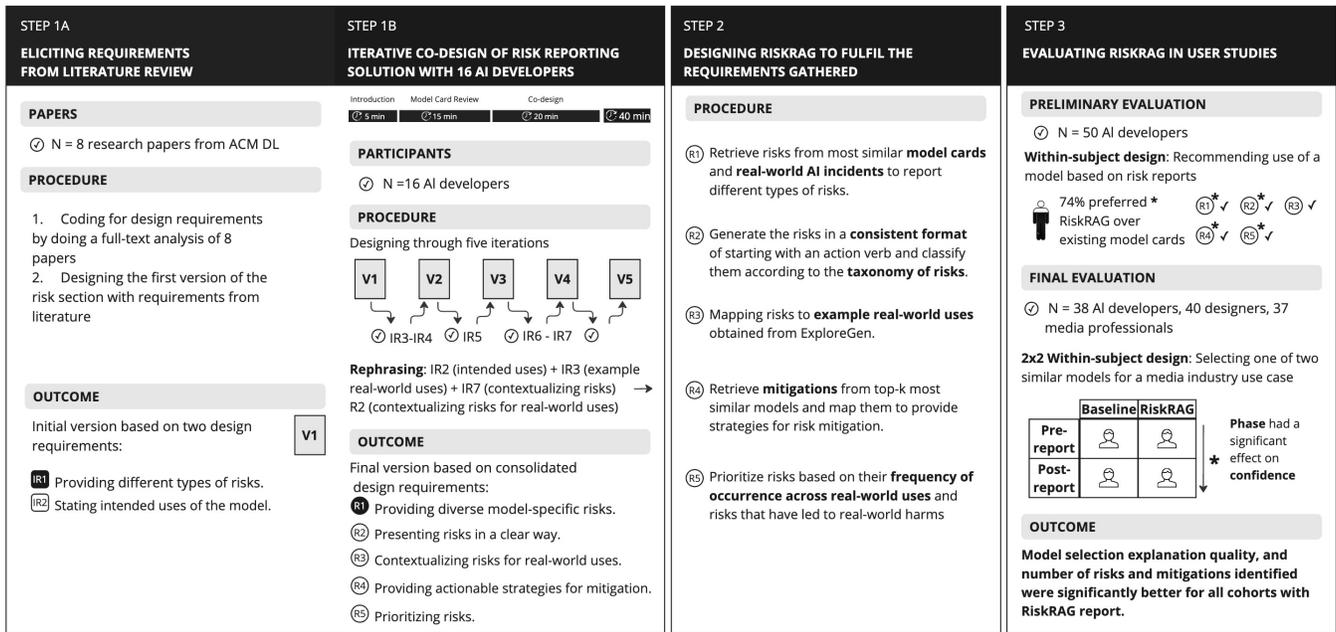


Figure 2: Our approach consists of three steps. *Step 1:* We identified design requirements through literature and a co-design study with AI developers (§3). Each co-design session included an introduction (~5 min), a model card review (~15 min), and a co-design task (~20 min). After five iterations, we finalized the key design requirements. *Step 2:* We developed RiskRAG, using retrieval-augmented generation to generate risk reports aligned with these design requirements, leveraging data from model cards, and incident reports (§4). *Step 3:* We evaluated RiskRAG reports in two user studies (§5). In the preliminary study, 50 AI developers compared a RiskRAG report to a baseline model card when assessing an AI model for a high-risk hiring scenario. In the final study, 38 AI developers, 40 UX designers, and 37 media professionals compared RiskRAG reports to baseline model cards when selecting between two similar AI models for media industry tasks.

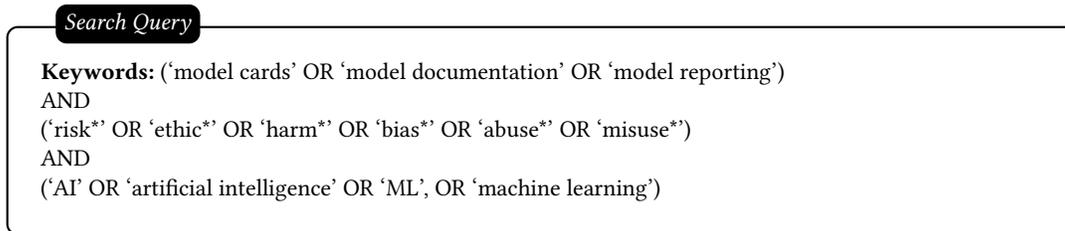


Figure 3: Search query used for the literature review within the ACM DL repository.

The codebook that we generated is provided in [Supp. Mat. 1.2](#). Finally, these led to two initial design requirements (IR) (Figure 2, Step 1A):

- IR1. Providing different types of model-specific risks.* Risk reporting of models should include different types of potential risks associated with model usage, including data and model limitations.
- IR2. Stating intended uses of the model.* Risk reporting of models should include intended uses of the model along with out-of-scope uses and misuse, as all these are related to risk reporting.

We used these requirements to generate the initial risk report, which was designed to guide AI developers through a step-by-step

process to evaluate their models' intended uses, assess model risks, and also identify potential gaps.

### 3.2 Design Requirements From Co-design

We conducted a series of one-on-one co-design sessions (Figure 2, Step 1B) with 16 AI developers (data scientists, researchers, and engineers), since they are the primary target users of our solution. Our co-design sessions were organized into five iterations. After each iteration, a refined risk report was developed based on user feedback. Consistent with prior research [6, 16], our sessions integrated semi-structured interviews with co-design activities to enhance the study's flow and efficiency. The resulting sequence of

**Table 1: Demographics of AI developers (P1-P16) who participated in our co-design sessions. Our participants have diverse expertise with practical, hands-on experience in deploying AI models across real-world applications.**

Version	ID	Gender	Age	Education	Expertise	Yrs of expr. in AI	Role	Hands-on Experience
V1	P1	Male	25	MSc	NLP	4	Data scientist	Chatbots for e-commerce applications
	P2	Male	24	MSc	NLP, CV	5	Researcher	Facial expression generation for interactive user applications
	P3	Female	22	PhD	NLP, CV	5	Researcher	Models for complex network analysis
V2	P4	Male	22	BSc	NLP, CV	5	Software engineer	Chatbots for hotel and finance applications
	P5	Male	28	MSc	CV	5	Researcher	Human face generation for virtual user applications
	P6	Male	33	PhD	Recommender systems	5	Lecturer	Recommender systems and data science
	P7	Male	32	PhD	CV, uncertainty quant	7	Researcher	Earth observation models for satellite data
V3	P8	Male	30	PhD	NLP	4	Researcher	Smart reply systems to enhance user interaction
	P9	Male	34	PhD	Privacy-preserving ML	10	Researcher	Augmented reality applications for preserving user privacy
	P10	Male	29	PhD	Reinforcement learning	6	Researcher	Modeling physical activities like running
	P11	Male	33	PhD	NLP, bayesian ML	5	Data scientist	Modeling marketing tasks and microscopic data
V4	P12	Male	40	PhD	Data science, ML engineering	2	Researcher	Anomaly detection in ECG signals for health monitoring
	P13	Female	31	PhD	ML security and privacy, NLP	7	Researcher	Diffusion models for AI safety
	P14	Male	22	BSc	CV	1	Data scientist	Misinformation detection in social media
V5	P15	Male	38	PhD	ML, generative AI	5	Data scientist	Personalization in retail for targeted customer offers
	P16	Female	26	PhD	On-device ML	5	Researcher	Model deployments on small devices like microcontrollers

risk report artifacts, generated after each of the five iterations, is presented in Appendix B, Figure 8.

**3.2.1 Goal.** The goal of these sessions was to understand AI developers’ requirements for reporting risks in model cards. The aim was also to develop a refined risk report at the end of the iterative co-design process that meets the identified requirements.

**3.2.2 Participants.** We aimed to achieve a diverse participant sample using snowball sampling, where the participants were asked to identify other potential subjects. We used the following screening criteria: (1) has graduate or undergraduate training in ML, statistics, or a related field, or has more than 2 years of experience with AI; (2) downloaded or uploaded a model from GitHub or HuggingFace during the past six months; and (3) age 18 or older. We recruited a total of 16 participants, including 13 men and 3 women, comprising an equal number of professionals from industry and academia. Participants brought diverse expertise, with extensive hands-on experience in developing and deploying AI models for a variety of real-world applications (see Table 1). Each iteration had three to four participants.

**3.2.3 Setup.** Before the session, we emailed participants with a demographic survey and a brief description of the session goals. They were also asked to provide us with a model card of a model they had used in the past six months. We prepared an initial version of a model card with only the intended uses and risk-related sections<sup>4</sup> based on initial requirements (Appendix Figure 8). We selected bert-base-uncased<sup>5</sup> from HuggingFace. This model is among the top 10 most downloaded models on the repository and the second most downloaded model to have a risk-related section.

**3.2.4 Procedure.** Each 40-minute session included three activities: **Introduction** (5 minutes): Participants introduced themselves, described their AI/ML projects, shared their experience with platforms

like GitHub and HuggingFace, and discussed the documentation practices of the models they used.

**A review of a model card** (15 minutes): Participants discussed the model card they brought, focusing on risk-related sections. They identified the sections they found most important for model selection, and evaluated the usefulness of the present risk-related content in anticipating potential model risks and challenges.

**A co-design task** (20 minutes): We introduced our version of the model card (based on the iteration) along with a task, to determine whether they would use the model for a specific high-risk use-case (i.e., a chatbot that answers questions about applicant resumes, and helps in filtering them) and to explain their reasoning. Participants were encouraged to suggest improvements for each model card section, focusing on information that would help them better complete the task, identify missing or inadequately represented details, and refine the risk section’s content and presentation. During earlier iterations, sessions emphasized understanding what participants needed to assess risks and justify model selection. In later iterations, more time was allocated to critiquing and co-designing the artifact itself.

Two authors facilitated each session: one led the questioning, while the other took detailed notes. Sessions were recorded, with participants’ consent, using online meeting software. After each session, we analyzed the key issues and updated the model card according to the requirements uncovered. This revised model card was then tested in subsequent co-design sessions with the next set of participants. By the time we developed the five versions of the model card (refer to Appendix Figure 8 for an overview of the iterations, which are described in [Supp. Mat. 2.1](#)), it became clear that our co-design efforts had yielded sufficient insights. No new significant issues were emerging, indicating that the design had reached saturation. Following established practices in cyclical action research [67], we decided to conclude the iterative process at this point (see Figure 1 for a quick overview, and the final version of the risk report is in [Supp. Mat. 4.2](#)).

**3.2.5 Gathering initial design requirements from participants.** After each iteration, two authors conducted a thematic analysis [9, 10] of the session’s transcripts, enriching them with session notes. We

<sup>4</sup>From the literature review, we found that risks of AI models were spread across different sections. Hence, we considered any of the following sections to be risk-related: intended uses, out-of-scope uses, risks, limitations, bias, ethical considerations, and responsibility and safety. For mitigations, we added recommendations subsection.

<sup>5</sup><https://huggingface.co/google-bert/bert-base-uncased>

employed an inductive coding approach where we coded the data to comprehend and highlight the requirements and issues raised by the participants regarding risk reporting. These codes were then jointly discussed and resolved for any disagreements. These were then arranged into relevant themes to derive a list of design requirements to be addressed in the next iteration of the model card (codebook used for each iteration is in [Supp. Mat. 2.2](#)). These co-design sessions with AI developers surfaced five additional design requirements, which led to the design of a model card artifact where no further refinements were deemed necessary (Figure 2, Step 1B):

- IR3. Providing example real-world uses of the model.* Risk reporting should include example real-world uses of the model as it can help users visualize how the model can be appropriately used and the potential risks associated with real-world scenarios. Note that this requirement differs from providing intended uses (*IR2*), which are more general (e.g., text generation tasks), while the participant asked for specific and concrete examples (e.g., the model can be used by journalists for generating news summaries). For example, P3 mentioned “I would see like if there are some examples of the applications. Like specific applications where it can potentially have the risks. Then it will be helpful.”, while P1 explained “It can give some practical world applications where it can be used. It just says a sequence classification...some real applications.”
- IR4. Providing strategies for mitigation.* Risk reporting should include recommendations or guidelines on how to mitigate the risks. For example, P4 expressed frustration with the original model card “... because it’s not fine-tuned for that specific task... even if they did give instructions [for use], it does not supply alternative solutions.”
- IR5. Presenting risks in a structured and easy-to-understand way.* Risks should be organized clearly and concisely, making them easy to comprehend and act upon. For instance, P7 commented “I think sort of a clear structure [would be helpful] few people will read through like a whole text section about this thing, unless they really dive into this topic... So somehow clearly structuring that... I guess it would be bullet points or like some diagram that has a quick summary of these are risks and biases and so on. And then more detailed information below... people have short attention spans.”
- IR6. Prioritizing risks.* Risks should be presented in a prioritized order, reflecting their impact and importance with which they need to be addressed. As P8 noted, “Potentially some way of just visually showing that, OK, these three out of the six have been highlighted as being major risks.”
- IR7. Contextualizing risks for specific real-world uses.* Risks should be clearly linked to particular real-world uses, making it easier to understand their relevance and impact in specific contexts. For instance, P3 expressed this requirement by saying “like when I ask about my application then if it can answer the possible risks and it will be really helpful.”

**3.2.6 Rephrasing the design requirements.** We consolidated the design requirements gathered from the literature and co-design process into five main requirements. We did this to make them more focused and orthogonal to each other, as, for example, some of the requirements were more specific versions of the other. This

also enhanced clarity, making it easier to implement the requirements in practice. For instance, the three initial requirements *IR2* (intended uses), *IR3* (example real-world uses), and *IR7* (contextualizing risks) were all consolidated into a single new requirement *R3* that encapsulates all of them.

Final design requirements:

- R1. Providing different types of model-specific risks.*
- R2. Presenting risks in a structured and easy-to-understand way.*
- R3. Contextualizing risks for specific real-world uses.*
- R4. Providing actionable strategies for mitigating risks.*
- R5. Prioritizing risks.*

## 4 Designing a Risk Reporting Solution (RiskRAG) Based on the Requirements

To meet the identified design requirements for AI model risk reporting (Figure 2, Step 2), we developed RiskRAG (Figure 4), a Retrieval Augmented Generation (RAG) based solution. An automated solution can effectively assist AI developers by simplifying the complex task of envisioning and documenting AI model risks, ensuring consistent and thorough reporting across different models. RAG is well-suited for this task because it combines retrieval-based and generation-based methods, ensuring that identified risks are relevant and grounded in real-world knowledge sources. By leveraging real-world and human-written datasets with risks, RiskRAG provides a reliable solution that complements developers’ expertise, making it easier to report risks that meet all design requirements gathered.

### 4.1 RiskRAG Architecture

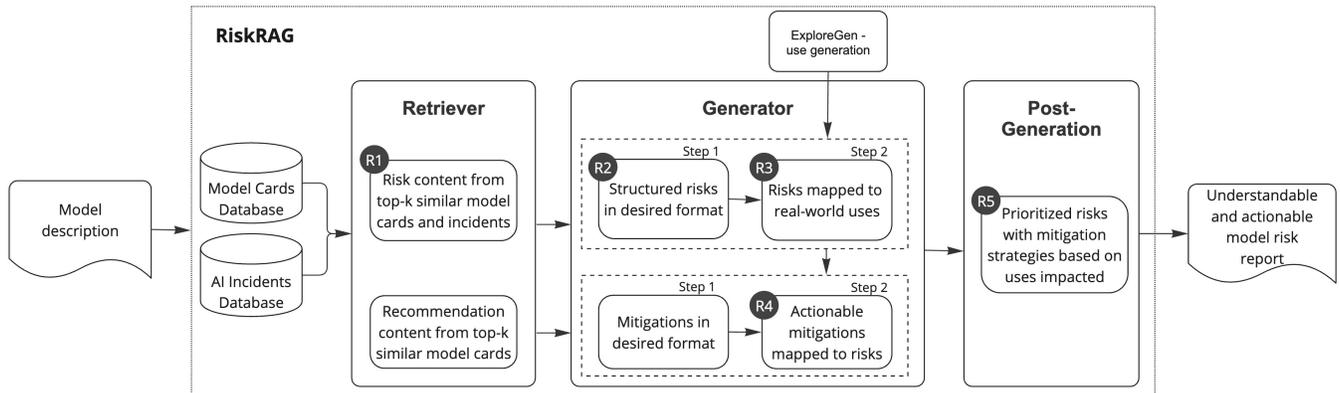
RiskRAG uses a standard RAG architecture [36] (naive RAG as described by [23]), which combines two stages: a retrieval model and a generation model. We used pre-trained models for both, as previous work [55] has shown that this method performs well without needing extra training and generally does better than fine-tuning models with specific data.

**4.1.1 Dataset.** We used two complementary datasets for retrieval: model cards (a source of both model risks and mitigation strategies) and the AI incidents database (a source of risks that resulted in real-world harms).

**Model cards dataset.** We downloaded a snapshot of the model repository published on HuggingFace<sup>6</sup> in July 2024 using the HF Hub API<sup>7</sup>. This consisted of 765,973 model repositories, out of which 461,181 (60%) had model cards. For each collected model card, we used regular expressions to search for risk-related sections. In particular, we searched for risks, limitations, bias, ethical considerations, out-of-scope uses, misuse, responsibility and safety sections. This led to 64,116 (14%) model cards with risk-related sections. In the absence of standardized and strict content requirements by HuggingFace, collected model cards were mostly incomplete, and many risk sections were only minimally modified copies of existing ones. Specifically, among the 64,116 model cards with risk-related sections, a huge majority (96%) had risk sections that were exact duplicates of another card. We further filtered out this dataset

<sup>6</sup><https://huggingface.co/models>

<sup>7</sup>[https://huggingface.co/docs/huggingface\\_hub/v0.5.1/en/package\\_reference/hf\\_api](https://huggingface.co/docs/huggingface_hub/v0.5.1/en/package_reference/hf_api)



**Figure 4: Architecture of RiskRAG.** We denote with *R1-R5* different steps aiming at fulfilling the design requirements from *R1* to *R5*. The input to RiskRAG is a description of the model for which a risk report needs to be generated. The retriever first extracts risk-related content from the top-*k* similar model cards and AI incidents (*R1*). The generator then adapts these risks into a standardized format and structures them using the risk taxonomy in [71] (*R2*). The ExploreGen LLM module [29] generates examples of real-world uses to which different risks are mapped to (*R3*). Mitigation strategies are similarly retrieved from model cards, formatted, and mapped to the corresponding risks (*R4*). Finally, risks are prioritized based on the number of uses they were mapped to and whether they have resulted in real-world incidents (*R5*).

and retained 2672 model cards having unique risk-related sections (we kept the most downloaded among the cards with duplicate sections), as our final model cards dataset (see Table 2 for statistics). **AI incidents dataset.** The AI Incident Database<sup>8</sup> is a publicly accessible resource that catalogues instances where AI systems have caused harm or failed in significant ways. By documenting these events, the database aims to promote transparency, improve understanding of AI risks, and guide the development of safer, more reliable AI systems. We chose to utilize this database as an additional source of risk information because it provides a crucial, real-world perspective on the various types of AI model risks that have manifested in deployments. As of March 2024, there were 649 incidents and 3412 reports, each incident derived from one or more reports. For example, one of the incident descriptions is “Meta’s open-source large language model, LLaMA, is allegedly being used to create graphic and explicit chatbots [...] that participate in text-based role-playing allegedly involving violent scenarios like rape and abuse.”<sup>9</sup> We collected the descriptions, metadata, and news reports about these 649 incidents as our AI incidents dataset.

**4.1.2 Retriever.** RAG retrievers are used in tasks like question answering, where the answers must be retrieved by comparing queries to source documents using cosine similarity. We used this method to retrieve risk-related sections from similar models as well as similar descriptions from incidents by treating the model description as the query. Our source documents included both model cards and AI incident descriptions. We computed contextual embeddings for query model descriptions and source documents. We calculated similarity scores between the query and source documents to identify the top-*k* most similar models and incident descriptions in each

dataset. Despite differences in presentation, contextual embeddings effectively capture semantic meaning across model cards and incident descriptions, as shown by prior research demonstrating their ability to handle complex linguistic structures, ambiguous word usage, and novel or domain-specific terms [2, 50, 58]. Some incidents specify model names, as the one in §4.1.1 involving the Llama model, which was matched to variants of Llama, as well as similar text generation models such as falcon-7b or phi-2. Other incidents lack specific model names but allow inference of the AI system’s capabilities, linking them to relevant model types. For example, the incident described as “alleged AI-generated photo alteration leads to inappropriate modifications in speaker’s conference picture”,<sup>10</sup> resulted from an Image-to-Image generation model and was associated with models such as flux-ip-adapter-v2 or instruct-pix2pix. We experimented with *k* = 5, 10, 15 to optimize for best results. From the top-*k* model descriptions, we took their corresponding risk-related sections. These top-*k* risk-related sections and retrieved incident descriptions are given as input to the generator.

An analysis of our model card dataset showed that mitigation strategies are either in a dedicated section like *Recommendations* or *Responsibility and Safety* or integrated within risk-related sections. Therefore, we used a combination of top-*k* retrieved risk-related and recommendation sections for extracting mitigation strategies.

We experimented with one sparse model: tfidf n-gram and three dense embedding models: SFR-Embedding-2\_R, Linq-Embed-Mistral and bge-large-en-v1.5.<sup>11</sup> The dense models were selected based on their high rankings in the Massive Text Embedding Benchmark [50] (MTEB) leaderboard as of July 2024. The first one

<sup>8</sup><https://incidentdatabase.ai/>

<sup>9</sup><https://incidentdatabase.ai/cite/578>

<sup>10</sup><https://incidentdatabase.ai/cite/820>

<sup>11</sup>[https://huggingface.co/Salesforce/SFR-Embedding-2\\_R](https://huggingface.co/Salesforce/SFR-Embedding-2_R), <https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral>, <https://huggingface.co/BAAl/bge-large-en-v1.5>

led in overall performance across 56 datasets, the second one excelled in retrieval tasks, and the third was the top performer with the fewest parameters. The `tfidf n-gram` model was included to compare the performance of traditional sparse representations against state-of-the-art dense embeddings. We used the `n-gram` range of 1-2.

**4.1.3 Generator.** RiskRAG uses GPT-4o [20] as generator (prompt used is in [Supp. Mat. 3.1](#)), as it is one of the leading LLMs for a variety of generation tasks [40], which also balances cost with efficiency. We devised a two-step generation for risks:

- (1) From the top- $k$  retrieved risk-related sections and incident descriptions, we generated risks in the desired format of `verb + object + [explanation]`, starting with an action verb. We generated zero or more unique risks from each retrieved risk-related section, and zero or more unique mitigation strategies from risk and mitigation-related sections. Further, we classified risks along two dimensions based on the taxonomy by Weidinger et al. [71]: where they occur (capability, human interaction, or systemic), and the type of harm they represent (e.g., representation and toxicity, misinformation, malicious use). Risks generated from incident descriptions were labeled as those that resulted in real-world harm.
- (2) RiskRAG uses ExploreGen [29] to generate a set of realistic and diverse model uses described using a five-component format: *domain*, *purpose*, *capability*, *AI deployer*, and *AI subject* [26]. ExploreGen outputs uses across 46 varied domains. We prompted it to additionally sort these uses by their likelihood and took the top four as examples. Each generated risk was mapped to a real-world use based on its relevance to the use.

We devised a similar two-step generation for mitigation strategies:

- (1) From the top- $k$  retrieved risk-related and recommendation-related sections, we generated one or more unique mitigation strategies in the same desired format as for risks.
- (2) Additionally, each generated mitigation strategy was mapped to one or more of the generated risks for which it was relevant.

After generating, RiskRAG prioritizes the risks (Figure 4, post-generation) based on how frequently they are mapped to example real-world uses, assuming that risks affecting more uses have a greater potential impact. Additionally, risks that have resulted in real-world harms in AI incident data are given a higher priority. While quantifying the impact and priority of risks remains an open research challenge [56, 57], this approach offers a simple and practical initial method for risk prioritization.

## 4.2 Meeting the Design Requirements

RiskRAG meets *R1* (different types of model-specific risks) because the retriever pulls risks from the most similar model cards, and similar models usually share comparable risks and limitations. For example, models trained on similar datasets or fine-tuned from the same parent model often exhibit similar biases, ethical issues, and fairness concerns. For instance, `bert-base-uncased`<sup>12</sup> and

`distilbert-base-uncased`<sup>13</sup> (later derived from the former) exhibit similar biases related to gender and race, as noted in their model cards. In the same vein, issues related to model interpretability and robustness often arise in models with similar architectures, regardless of their specific application areas. By retrieving between  $k = 5$  and  $k = 15$  similar model cards, the retriever enables us to capture a substantial portion of the model-specific risks associated with the target model, which are predominantly technical and model-capability-related [58]. To capture a broader range of risks, especially human-interaction ones [68], RiskRAG also retrieves risks from real-world AI incidents linked to the model’s use. For example, the ChatGPT model is associated with AI incident 642, which describes a glitch that disrupted user interactions with nonsensical outputs.<sup>14</sup> Since not all incidents specify the underlying AI model, we link incidents to models performing the same task (e.g., incident 642 would also be linked to other similar text generation models). These incidents often reveal harms from human-AI interactions [68] that are underrepresented in model cards, further helping to meet *R1*. Furthermore, the generator adapts all identified risks, both those originating from similar models and those from related incidents, to the unique context of the target model. This adaptation process may involve dropping risks that are not applicable to the target model or modifying them to reflect the model’s specific characteristics. For example, a risk such as “underrepresents cultures using non-English languages” might be adapted to “underrepresents cultures using non-Chinese languages, if the target model is trained on Chinese rather than English text.”

RiskRAG meets *R2* (structured and easy-to-understand risks) by generating the risks in a consistent and actionable format (as described in §4.1). An example risk assigned to text generation models is: “undermines user trust by providing inappropriate suggestions.” This well-defined format ensures that risks are articulated clearly, minimizing ambiguity. When risks are framed with action in mind, it becomes easier to implement effective mitigation strategies. RiskRAG also structures these risks using the risk taxonomy [71] further helping to meet *R2*. For instance, the example risk above is classified under the category of *information & safety harms* and in the *human-interaction* layer.

RiskRAG meets *R3* (contextualizing risks for specific uses) by mapping each risk to example uses generated by ExploreGen, contextualizing it for these specific real-world applications. For example, if the model in question is designed for text generation, then the risk “undermines user trust by providing inappropriate suggestions” is applicable for the use *detecting harmful content*, while its risk “violates fairness in recruitment by giving false positive results” is applicable to the use *enhancing job matching*.

RiskRAG meets *R4* (actionable mitigation strategies) by retrieving mitigations from most similar model cards and mapping them to specific risks to provide strategies for risk mitigation. An example mitigation for the risk mentioned above is: “filter the outputs of the model for irrelevant or inappropriate suggestions.”

At last, RiskRAG meets *R5* (prioritizing risks) by leveraging risk frequency of occurrence across real-world uses and giving a higher importance to the risks that have led to real-world harms based

<sup>12</sup><https://huggingface.co/google-bert/bert-base-uncased>

<sup>13</sup><https://huggingface.co/distilbert/distilbert-base-uncased>

<sup>14</sup><https://incidentdatabase.ai/cite/642>

**Table 2: Statistics of our evaluation dataset for assessing RiskRAG. From the whole dataset of 2672 model cards that do not have risk content copied from each other, we took the top 10% most downloaded ones as our evaluation set.**

	Evaluation set	Remaining set	Whole dataset
Number of model cards	267	2405	2672
Average number of downloads	1181494.73	497.04	118508.41
Length of risk-related sections (characters)	1233.22	707.12	759.69

on the AI incident data. For example, “violates privacy rights by disclosing sensitive personal data” is given higher priority compared to “replicates inherent biases in data” as the former was retrieved from AI incidents and resulted in real-world harm.

To sum up, retrieving the top 5 to 10 similar model cards enables the retriever to capture the broader context surrounding the target model, while the generator refines this content to address the model’s unique characteristics and nuanced, context-specific risks.

## 5 Evaluating Our Risk Reporting Solution

Before evaluating whether RiskRAG produces risk reports that meet the design requirements, we first needed to determine its preliminary effectiveness in generating relevant risk content for these reports. To do so, we initially conducted a baseline evaluation (§5.1) of RiskRAG-generated content. Once we ascertained that our method performs well, in a preliminary user study (§5.2), we assessed the alignment of RiskRAG reports with the design requirements. In a final user study (§5.3), we evaluated whether RiskRAG can assist in selecting the most suitable AI model for a given use and support decision-making.

### 5.1 Baseline RiskRAG Evaluation

There are no standardized approaches for evaluating the RAG-generated content [66]. The challenges arise from variations in retrieved content and the common absence of ground truth for customized generation pipelines [23, 46]. In our case, there is also a lack of ground truth, as the model risk sections formatted according to all our requirements do not yet exist (e.g., risks based on uses sorted by priority). To address this challenge, we focused on evaluating the risk content format that most closely matches those available in existing model cards (i.e., individual risk/mitigation descriptions).

**5.1.1 Goal.** The main goal of the baseline evaluation was to establish that it produces relevant risk content from which the final risk report, meeting the design requirements, can be produced. An additional goal was to determine the optimal parameters for RiskRAG, i.e., the retrieval (embedding) model to be used in the retriever, and the number of similar model cards ( $k$ ) to be retrieved.

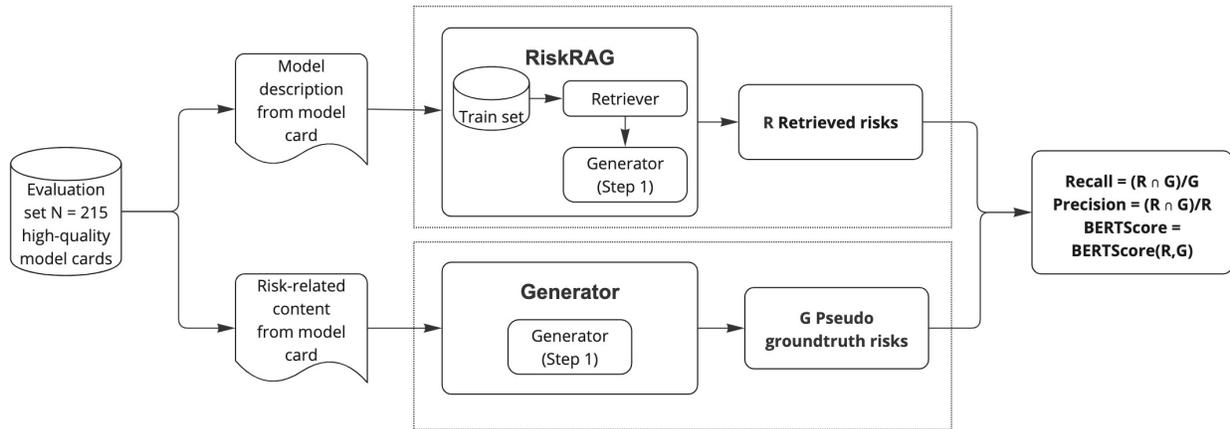
**5.1.2 Evaluation setup.** One more challenge for our evaluation was that the risk sections in existing model cards are often incomplete or sparse, making traditional evaluation methods with train and test data splits difficult (e.g., we cannot include many model cards with sparse risk sections in either the training or test set.). To address this, we created an evaluation set using high-quality model cards

as a pseudo ground truth. Specifically, we automatically extracted 267 model cards, i.e., the top 10% most downloaded ones (see “Evaluation set” in Table 2). We used popularity as a proxy for model card quality [37]. To ensure that these cards are indeed of high quality, we manually inspected a subset of 30 cards. Table 2 shows that, on average, these cards have longer risk reports compared to other cards, and our manual inspection confirmed that they contain non-sparse and generally carefully written risk content. Since RiskRAG generates individual risks rather than complete sections, we needed to adjust the evaluation set to match our output format. We accomplished this by processing all the risk sections from the evaluation set through step 1 of the RiskRAG generator (Figure 4). This resulted in a set of 215 model cards containing individual risks (even among the most popular model cards, some were found without risk content, and we dropped those), creating a pseudo ground truth ( $G$ ) in our proposed format. Finally, we compared different retrieval (embedding) models and tested various values of the parameter  $k$  to identify the best performer and determine the optimal  $k$ . The complete evaluation setup is shown in Figure 5. To determine the hyperparameters of the system, we also compiled a separate validation set by randomly sampling model cards ( $n = 25$ ) using the same process, selecting from those that were downloaded in the top 10% to top 20% range.

**5.1.3 Metrics.** We evaluated RiskRAG using two types of metrics. First, BERTScore [75] measures the overall similarity between the retrieved risks ( $R$ ) and the pseudo ground truth ones ( $G$ ). Second, precision and recall assess the efficiency of retrieving individual risks from the pseudo ground truth. To calculate precision and recall, we matched each retrieved risk ( $\in R$ ) with the pseudo ground truth risks ( $\in G$ ) using BERTScore rather than direct string matching. Direct string matching often misses contextually similar risks with different phrasing, while BERTScore, by leveraging contextual embeddings, more accurately assesses alignment between retrieved risks and pseudo ground truth. We considered a retrieved risk as correctly matched, if its BERTScore with a pseudo ground truth risk exceeded a threshold of 0.6. This threshold was determined through manual annotation on the separate validation set of randomly sampled pairs of risks (retrieved, and pseudo ground truth). For each pair, two authors evaluated whether the retrieved risk was contextually relevant to the pseudo ground truth risk. We then selected the threshold value that best aligned BERTScore matches with these manual annotations.

Recall is calculated as the ratio of correctly retrieved risks to the total number of pseudo ground truth risks ( $R \cap G$ )/ $G$ , while precision is the ratio of correctly retrieved risks to the total number of retrieved risks ( $R \cap G$ )/ $R$ . We used the original implementation of BERTScore with contextual embeddings from a pre-trained language model `DistilBERT` [62].

**5.1.4 Results.** Our evaluation results are presented in Table 3. While there are no previous studies directly targeting our specific task, the closest approach for comparison is CardGen [39]. CardGen leverages RAG to generate model card sections based on input from related papers and GitHub repositories. The BERTScores reported by CardGen for risk-related sections, using various embedding models, range from 0.53 to 0.59, which is comparable to our results (0.53 for top- $k = 5$  and `tfidf n-gram`). It is important to note that our



**Figure 5: Baseline RiskRAG evaluation.** This evaluation is performed on the evaluation set consisting of the top 10% most downloaded model cards (Table 2). We first produced risks  $R$  with RiskRAG for each of these cards using only their model descriptions. To assess the quality of RiskRAG’s output against the existing risk sections of these cards, we parsed these risk sections through the generator (step 1) generating pseudo ground truth  $G$  to make them compatible with the risk content generated by RiskRAG, enabling direct comparison.

**Table 3: Results of the baseline RiskRAG evaluation.** Information retrieval (precision and recall), and the text generation (BERTScore) metrics are shown. The parameter  $top-k$  represents the number of most similar model cards from which risks were retrieved.

	$top-k = 5$			$top-k = 10$			$top-k = 15$		
	Precision	Recall	BERTScore	Precision	Recall	BERTScore	Precision	Recall	BERTScore
Linq-Embed-Mistral	0.32	0.71	0.51	0.20	0.75	0.42	0.15	<b>0.78</b>	0.37
SFR-Embedding-2_R	0.32	0.69	0.51	0.20	0.75	0.42	0.14	0.76	0.36
bge-large-en-v1.5	0.27	0.60	0.48	0.18	0.66	0.40	0.13	0.68	0.35
tfidf n-gram	<b>0.34</b>	0.71	<b>0.53</b>	0.21	0.75	0.42	0.16	0.77	0.37

task, which requires outputting risk sections in a format different from the original model cards, adds an additional layer of complexity to the retrieval process, making it more challenging compared to CardGen.

Table 3 reports precision and recall along with BERTScore. Higher recall indicates that RiskRAG successfully retrieves more risks present in the pseudo ground truth model card, while higher precision reflects fewer false positives, meaning more of the retrieved risks are indeed part of the pseudo ground truth. For  $top-k = 5$ , *tfidf n-gram* achieved the highest precision (0.34) and BERTScore (0.53). Both *Linq-Embed-Mistral* and *SFR-Embedding-2\_R* also performed well, each with a precision of 0.32 and a BERTScore of 0.51, demonstrating their competitive accuracy and similarity. As the  $top-k$  value increases, all three models maintained strong recall, with *Linq-Embed-Mistral* slightly ahead, reaching a recall of 0.78 at  $top-k = 15$ . This suggests that while *tfidf n-gram* excels in precision and BERTScore, *Linq-Embed-Mistral* is slightly better at retrieving a broader set of relevant risks. Precision is comparatively low, and declines as  $top-k$  increases. However, since the risk sections in model cards in our evaluation set are likely incomplete, lower precision does not always indicate false positives. Many retrieved risks may be missing from the pseudo ground truth but remain relevant to the model as they are sourced from similar models. Additionally, Step 2 of the generator (not included in this part

of the evaluation) can filter out irrelevant risks, refining the results further. Therefore, precision is less critical in our evaluation than recall or BERTScore.

To confirm that RiskRAG generates a broad and relevant set of risks despite low precision, and to resolve the tie between *tfidf n-gram* and *Linq-Embed-Mistral*, we qualitatively examined the risks for the ten most downloaded models from our evaluation set. We chose  $k = 10$  for its broader risk coverage over  $k = 5$ , while avoiding the lengthy lists seen with  $k = 15$ . Indeed, an initial manual analysis on a validation set showed that  $k = 10$  best balanced coverage and conciseness, given the incomplete risk sections in many model cards. Upon the manual evaluation of the ten most downloaded cards, we found that most of the risks retrieved were relevant to the model in question, and were indeed missing in the original card. For instance, the card for *google/flan-t5-large*<sup>15</sup>, a multilingual text generation model, listed risks such as biases in training data, and harmful, inappropriate, or explicit content generation. RiskRAG expanded this by identifying additional relevant risks, including hallucination, toxicity, misinformation, unsupported languages, malicious use, and representational harms, such as racial and gender stereotypes in online data. Although these risks

<sup>15</sup><https://huggingface.co/google/flan-t5-large>

were not documented in the original card, they are relevant for comprehensive risk assessment. We also observed that low precision was partly due to similarly worded risks across top-10 results and the presence of use-specific risks from similar models that did not directly pertain to the model for which risks were retrieved. These issues get mitigated in Step 2 of the generator, where irrelevant risks are dropped, and those specific ones are adapted.

We chose Linq-Embed-Mistral for the later user evaluation (§5.2) for two reasons: (1) Comprehensive coverage. Linq-Embed-Mistral provided a broader and more detailed set of risks. For DistilBERT, tfidf n-gram primarily highlighted biases such as “produces biased predictions despite neutral training data” and “transfers bias to all fine-tuned versions”. Linq-Embed-Mistral included additional risks related to domain-specific performance and language issues like “underperforms on text from different domains” and “underperforms on non-English languages”. (2) Relevance. Linq-Embed-Mistral identified risks more pertinent to specific tasks. For example, it captured risks related to image classification for the OpenAI CLIP<sup>16</sup> model, such as “reduces performance when input images are resized” and “memorizes duplicated images in the training data”, which tfidf n-gram missed. Overall, while tfidf n-gram highlighted major risks, Linq-Embed-Mistral offered a more broad and relevant assessment, making it the better choice for detailed risk reporting.

## 5.2 Preliminary User Study

We conducted a preliminary user study (Figure 2, Step 3 top) to understand how AI developers perceive RiskRAG’s risk reports relative to existing model card risk reports.

**5.2.1 Goal.** To assess whether RiskRAG’s reports meet the identified design requirements and are preferred over standard model card risk reports when deciding whether to use an AI model for a given task.

**5.2.2 Study design.** We conducted a within-subject study where each participant evaluated two versions of a model risk report: a baseline version containing the risk-related sections of the original model card (control), and the model risk report generated by RiskRAG (§4) (treatment). Specifically, we chose two similar text generation models: Phi-3-mini-128k-instruct<sup>17</sup> (downloaded over 200K times in July 2024) and StableBeluga2<sup>18</sup> (downloaded over 130K times in the same period). These models were selected because their original model cards feature relatively rich risk sections, allowing for a fair comparison of the baseline with the RiskRAG report. Participants were then asked to consider one of two high-risk (according to the EU AI Act) uses for the models: (1) *a chatbot to evaluate and rank pre-interview assessments* and (2) *a chatbot that answers questions about applicant resumes and helps in filtering them*. We focused on high-risk scenarios because effective risk reporting is crucial in such contexts. Additionally, selecting applicants for a team is commonly experienced by developers, making these model uses relatable and relevant for our participants.

We developed a web-based survey that included a real-world task to be performed in both control and treatment conditions:

*“Write a short email to your line manager asking for approval to use this model. In your email, explain the technical and ethical reasons why this model should be used, but also be candid about any potential risks and discuss how they can be mitigated.”*

The vast availability of AI models today makes it realistic for developers to argue for the use of any particular one, and discuss trade-offs with management. In the post-COVID world, where a significant portion of work communication occurs online, as many people still choose to work remotely [59], sending an email was also considered a practical and relevant task.

**5.2.3 Metrics.** To assess how effectively RiskRAG met the requirements compared to the baseline risk sections from the original model cards, we measured the following (left panel of Figure 6):

*Q1. Does the risk report provide reliable information and cover a wide range of risks? (R1)*

We measured this by adapting an item for *completeness* and another for *perceived accuracy* from the AIMQ information quality assessment scale [35].

*Q2. Does the risk report explain the risks in a clear and concise manner that is easy to understand? (R2)*

We used items on *understandability* and *concise representation* from the AIMQ scale [35] to measure this.

*Q3. Is the content of the risk report relevant to the use case presented in the task? (R3)*

We adapted and measured an item on information *relevance* from the AIMQ scale [35].

*Q4. Does the risk report offer clear strategies for mitigating risks? (R4)*

We adapted an item on *ease of operation* from the AIMQ scale [35] to measure this.

*Q5. Does the risk report effectively prioritize the risks? (R5)*

We developed a custom item to measure this specific requirement that emerged during our co-design sessions.

Each question was rated on the scale from 1 (strongly disagree) to 5 (strongly agree).

The last metric we measured was the *preference* towards the baseline risk report or treatment risk report from RiskRAG. For uniformity, all metrics were measured using a 5-point Likert scale. Finally, we also had two open-ended questions: “In what ways did the report succeed in assisting you in completing the task?” and “In what ways did the report fall short of assisting you in completing the task?”

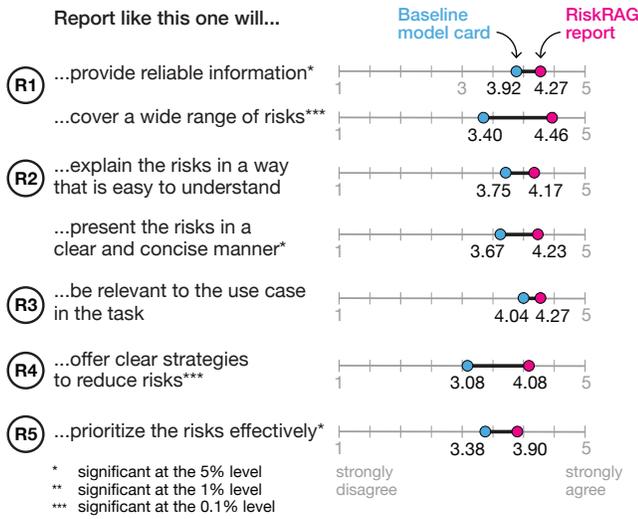
**5.2.4 Participants.** We focused on AI developers, who are responsible for assessing how AI systems perform in specific use cases, including evaluating human interaction effects and technical capabilities within their applications [71]. We recruited 50 AI developers through the online recruitment platform Prolific.<sup>19</sup> To ensure the suitability of participants, we applied four a priori inclusion criteria, targeting individuals who held individual contributor roles, worked in an engineering function within their organization, were employed specifically for coding tasks, and used AI multiple times per week. Additionally, we used two items from the AI literacy

<sup>16</sup><https://huggingface.co/openai/clip-vit-large-patch14>

<sup>17</sup><https://huggingface.co/microsoft/Phi-3-mini-128k-instruct>

<sup>18</sup><https://huggingface.co/petals-team/StableBeluga2>

<sup>19</sup>See <https://www.prolific.com>, last accessed Aug 2024.



**Figure 6: Quantitative results from the preliminary user study: RiskRAG report outperformed baseline model cards across all the metrics. We had seven questions mapping to our design requirements to which participants were asked to answer on a Likert scale from 1 (“strongly disagree”) to 5 (“strongly agree”). RiskRAG significantly outperformed the baseline model cards, with a one-point higher rating, moving from slight agreement to clear agreement on risk coverage, and from neutral to agreement on mitigation clarity.**

scale [12] to control for participants’ AI literacy and their attitudes towards AI.

Participants’ ages ranged from 18 to 29 (40%) and 30 to 39 (60%). All participants were male, residing in the U.S., and working as individual contributors in non-managerial technical roles, including engineering (33%), design (17%), data analysis (17%), and research and education (17%). Regarding ethnicity, 40% identified as Asian, 20% as Black, 20% as Mixed, and 20% as White. In accordance with our study requirements, 60% of participants reported using AI in their work multiple times a week, while 40% used it daily.

**5.2.5 Procedure.** The entire procedure took place in three steps. In the first step, participants answered a demographics questionnaire. In the second step, participants were provided a brief introduction to the tasks, followed by the first task in which participants had to read either the control or treatment risk report, and write the email. In the third step, they answered the questions about the chosen metrics. They then proceeded to view the other risk report and answered the same set of questions.

To counterbalance potential order effects, the sequence in which the baseline and treatment risk reports were shown to participants was randomized. To eliminate any effects from the type of AI models shown, each participant reviewed risk reports of two different models with different real-world uses, randomly assigned to ensure a thorough evaluation. Baseline and RiskRAG reports for the two models are in [Supp. Mat. 4](#).

**Table 4: Statistical significance of the results in the preliminary user study: Wilcoxon signed-rank test results between the baseline model cards and the RiskRAG report. The test showed statistical significance across the majority of the chosen quantitative metrics, indicating that the RiskRAG report outperformed the baseline across all the identified requirements: \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$ ; ns: $p > 0.05$ .**

metric	z-statistic	p-value
provide reliable information (R1)	63	*
covers a wide range of risks (R1)	60	***
explain risks in a way that is easy to understand (R2)	100.5	ns
present risks in a clear and concise manner (R2)	103	*
be relevant to the use case in the task (R3)	87	ns
offer clear strategies to reduce risks (R4)	77	***
prioritize the risks effectively (R5)	149	*
preference between RiskRAG report and baseline	331.5	**

**5.2.6 Analysis.** After confirming non-normality with the Shapiro-Wilk test, we applied the Wilcoxon signed-rank test to assess if the RiskRAG enhancements produced statistically significant improvements over the baseline. The results of the thematic analysis of qualitative feedback from open-ended questions are detailed in Appendix D.

**5.2.7 Results.** Figure 6 presents the breakdown of the quantitative results. Across the questions, the participants rated the RiskRAG report more favorably than the baseline model card. The largest difference is evident for the questions on offering clear mitigation strategies (R4), where the RiskRAG report scored 4.08 (“agree”), and baseline card 3.08 (“neutral”), and for covering a wide range of risks (R1), where the report scored 4.46 (“clearly agree”), and the card 3.40 (“slightly agree”). While participants gave higher scores to the RiskRAG report for explaining risks in easy to understand way (R2) (4.17 for our report versus 3.75 for the baseline) and being relevant to the use case in the task (R3) (4.27 vs. 4.04), the statistical tests did not show statistical significance on these questions. The differences in all the other questions, i.e., providing reliable information, presenting risks in a clear and concise manner, and prioritizing them effectively, were statistically significant (Table 4). The preference for the RiskRAG report over the baseline was statistically significant, with 74% favoring it.

## 5.3 Final User Study

Through the preliminary user study, we established that RiskRAG generates risk reports that fulfil all identified design requirements and are preferred over existing risk reports in model cards. Subsequently, we conducted the final user study (Figure 2, Step 3 bottom) to assess its effectiveness in decision-making and actionability.

**5.3.1 Goal.** To assess whether RiskRAG reports improve understanding of model risks, facilitate critical evaluation, and support envisioning mitigation strategies better than standard model card risk reports.

**5.3.2 Study design.** This study employed a 2×2 within-subject design. The task was to assess two models for a real-world use case, and select the one deemed more suitable. Each participant completed this task across two conditions:

- (1) **Control:** risk-related sections of the original model card; and
- (2) **Treatment:** RiskRAG-generated risk report.

For both conditions, the task was divided into two *phases*:

- (1) **Pre-report phase.** Participants chose a model after reviewing descriptions of two models and their intended use, providing an explanation for their choice.
- (2) **Post-report phase.** Participants reviewed a brief incident tied to the intended use and the risk reports (original model card for control, RiskRAG for treatment). They could reconsider and change their previous model choice, explaining in either case their final decision. This step assessed the impact of the risk report on their decision-making.

This design enabled a direct comparison of participants' decision-making with and without RiskRAG reports, as well as with and without baseline model card risk sections, providing insights into their different impacts on model selection.

To test RiskRAG's ability to generalize beyond text generation models that we used in the preliminary study, we selected two pairs of models, both of different type than text generation, each coupled with an appropriate real-world use and an associated incident:

- (1) Multimodal Models (Image or Text-to-Text): `idefics-80b-instruct`<sup>20</sup> and `paligemma-3b-mix-448`<sup>21</sup>, which take both image and text inputs and produce text outputs. Use: Developing a system for a media organization to identify people and objects in photos and generate alternative text descriptions for web pages. Incident: The model mistakenly labelled a Black couple as "gorillas".<sup>22</sup>
- (2) Automatic Speech Recognition (ASR) Models (Speech-to-Text): `whisper-large-v3a`<sup>23</sup> and `canary-1b`<sup>24</sup>, which process speech input and convert it to text. Use: Creating a system to transcribe spoken content into text for broadcast subtitles. Incident: The model hallucinated violent language and fabricated details, particularly during extended pauses in speech.<sup>25</sup>

These medium- to high-popularity models, with rich risk-related sections in their original cards, provided a strong baseline for RiskRAG. The models were selected to have comparable strengths and risks, ensuring no definitive "right" choice between them. This allowed us to focus on how participants deliberated between different pairs of models.

**5.3.3 Metrics.** We measured the *explanation quality* to assess differences between the baseline condition and the treatment condition in the post-report phase. Three authors, with extensive expertise in responsible AI, human-computer interaction, computer vision, and NLP, and with a strong publication record in AI applications, risk reporting, impact assessments, and user study design, independently annotated the explanations according to a scoring rubric. Prior to performing any annotations, all evaluators participated in a calibration session to ensure a consistent understanding and

application of the scoring rubric, which is detailed in the [Supp. Mat. 5.1](#), and summarized here:

- (1) *Number of identified risks:* Evaluators counted the number of identified risks in the explanation. These risks were either in the context of the real-world use or supported model selection, demonstrating the understanding of the report's content.
- (2) *Number of proposed mitigations:* Evaluators counted the number of appropriate mitigation strategies proposed to counter the identified risks.
- (3) *Task quality:* Evaluators rated how effectively the explanation communicated the trade-offs between risks and benefits for the selected model. This was scored on a scale from 1 to 5, where a score of 1 reflected vague reasoning, lack of argumentation, or no clear call to action, and a score of 5 reflected strong alignment with the task, robust mitigation strategies, and a well-structured argument featuring diverse trade-offs.

The inter-annotator agreement, calculated by a Fleiss' kappa, ranged from 0.82 for the number of identified risks to 0.78 for the task quality. This confirmed that the evaluators agreed with each other strongly. For each metric listed above, the final score was determined by averaging the scores given by three evaluators. The *overall explanation quality* was then calculated as the average of the final scores for all three metrics.

We measured the following *decision metrics* in pre- and post-report phases for both conditions:

- (1) *Decision confidence:* Participants' self-reported confidence in their model choice, measured with the question: 'How confident are you in your ability to choose the most appropriate AI model for this task?'. This was measured using a 5-point numerical scale.
- (2) *Decision time:* The time participants took to make their model selection.
- (3) *Preference:* We also measured the *preference* between reports, as used in the preliminary study (§5.2.3)

**5.3.4 Participants.** We conducted the study in three cohorts (Table 5): 38 AI developers, 40 UX designers and 37 media professionals recruited through Prolific. For the 2×2 within-subject study, we conducted a priori power analysis. A repeated measures ANOVA with medium effect size ( $f = 0.40$ ,  $\alpha = 0.05$ ,  $1 - \beta = 0.95$ ) indicated a minimum requirement of 15 participants, which was met and exceeded by all our cohorts. In the first cohort, we recruited developers using the same inclusion criteria as in §5.2.4. In the second cohort of UX designers, we recruited individuals who held individual contributor roles, worked in a design or creative function within their organization and used AI at least once a week. The third cohort included professionals in journalism, marketing, communications, design, and creative roles, who use AI at least once a week. We selected this group because they are the primary users of the applications featured in the task. AI developers were the most knowledgeable in the task, technology, and AI across the three cohorts, and the task was most similar to their day-to-day tasks at work.

<sup>20</sup><https://huggingface.co/HuggingFaceM4/idefics-80b-instruct>, last accessed Nov 2024.

<sup>21</sup><https://huggingface.co/google/paligemma-3b-mix-448>, last accessed Nov 2024.

<sup>22</sup><https://incidentdatabase.ai/cite/16/>, last accessed Nov 2024.

<sup>23</sup><https://huggingface.co/openai/whisper-large-v3>, last accessed Nov 2024.

<sup>24</sup><https://huggingface.co/nvidia/canary-1b>, last accessed Nov 2024.

<sup>25</sup><https://incidentdatabase.ai/cite/732/>, last accessed Nov 2024.

**Table 5: Self-reported knowledge and demographic characteristics of participants in the final user study.**

Control	Characteristic	AI Developers (n=38)	UX Designers (n=40)	Media Professionals (n=37)
Expertise	Task	3.97 $\pm$ 0.75	3.60 $\pm$ 1.08	3.65 $\pm$ 1.01
	Technology in general	4.34 $\pm$ 0.67	4.15 $\pm$ 0.58	4.05 $\pm$ 0.74
	Artificial Intelligence	4.13 $\pm$ 0.74	3.88 $\pm$ 0.91	3.86 $\pm$ 0.75
Task similarity	Similarity with day-to-day tasks	3.34 $\pm$ 0.94	2.92 $\pm$ 0.94	2.92 $\pm$ 1.04
Age	18-29 years	47.4%	45.0%	40.5%
	30-39 years	26.3%	22.5%	35.1%
	40-49 years	18.4%	22.5%	10.8%
	50-59 years	5.3%	7.5%	10.8%
	60 years and above	2.6%	2.5%	2.7%
Sex	Female	26.3%	32.5%	48.7%
	Male	73.7%	67.5%	48.6%
	Prefer not to say	0	0	2.7%
Ethnicity	White	34.2%	47.5%	40.5%
	Black	34.2%	35%	35.1%
	Mixed	18.4%	12.5%	21.6%
	Asian	7.9%	5%	0
	Other	0	0	2.7%
	Not specified	5.3%	0	0

**5.3.5 Procedure.** The procedure consisted of three main steps. (1) Participants provided informed consent; (2) Participants received instructions for pre-report phase, selected a model, explained their choice, and reported their decision confidence; and (3) Participants received instructions for post-report phase, reviewed the risk report, reconsidered their selection, explained their final choice, and reported their confidence again. This process was repeated with a second set of models using the alternate risk report. To counterbalance order effects, the order of the baseline report and treatment report was randomized, and the two multimodal and two audio models were randomly assigned to either condition. Baseline and RiskRAG reports for all four models are in [Supp. Mat. 5.2](#).

**5.3.6 Analysis.** We used the Wilcoxon signed-rank test to analyse the difference in explanation quality between control and treatment. A 2x2 repeated measures ANOVA examined the effects on decision metrics, identifying main effects and interactions to assess whether RiskRAG reports significantly influenced confidence and time compared to the control. We conducted inductive thematic analysis [9, 10] to derive themes from participants' model choices and preference explanations. Following established coding procedures [47, 60], two authors coded the data for the baseline condition and report condition. The process involved familiarizing with the data, iterative coding with refinement, and resolving disagreements through discussion. This analysis yielded 62 codes organized into a thematic map with 11 sub-themes and 5 themes for the report condition, and 32 codes grouped into 6 sub-themes and 3 themes for the baseline condition. The codebook is provided in [Supp. Mat. 5.3](#).

**5.3.7 Results.** *Explanation quality* was higher for RiskRAG reports compared to the baseline reports across all the three cohorts (Figure

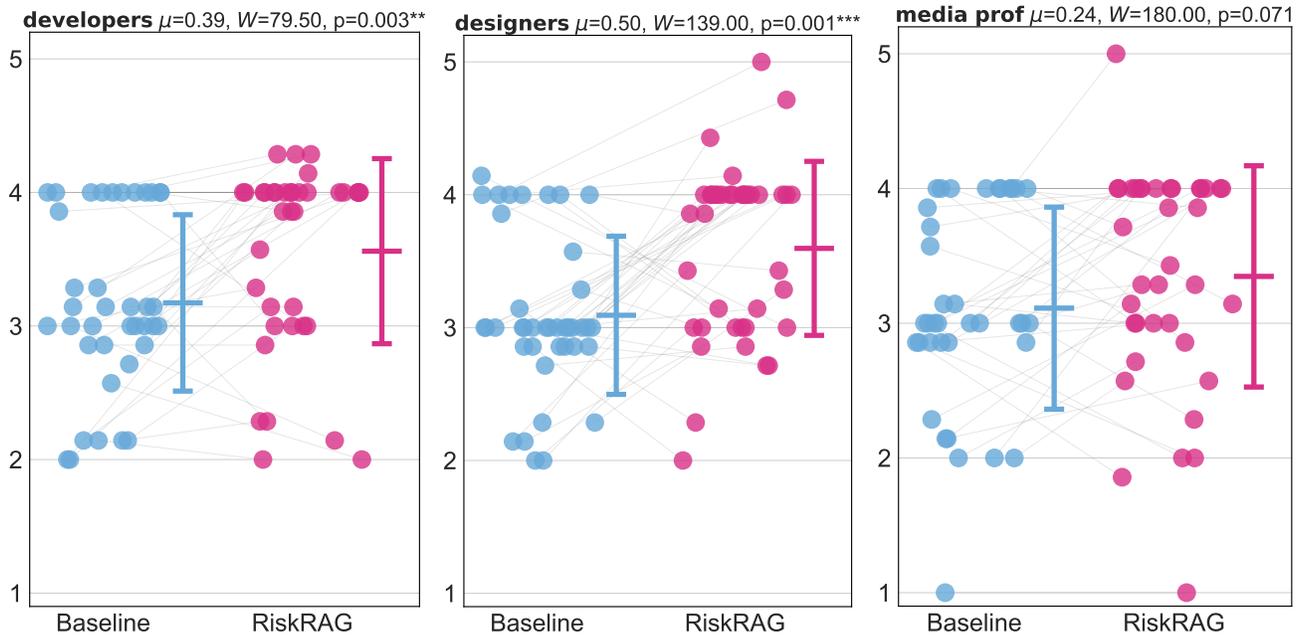
7). For developers, both the number of proposed risks and mitigations ( $W = 104, p = .041$  and  $W = 130, p = .021$ ), as well as task quality ( $W = 108, p = .011$ ) were significantly higher. For designers, the number of identified risks ( $W = 157, p = .016$ ) and task quality ( $W = 139, p = .001$ ) were significantly higher. Although the differences did not reach statistical significance, the explanations of media professionals were also of higher quality.

Regarding decision metrics, Table 6 shows that *decision confidence* was significantly different in post-report phase when compared to pre-report phase for the developers' ( $F(1, 38) = 6.11, p = .018$ ) and media professionals' ( $F(1, 38) = 6.11, p = .030$ ) cohorts. Their confidence decreased after interacting with the reports, with a greater drop observed in the treatment condition compared to the control. Designers showed a similar trend, although it was not statistically significant. We will later examine these lower confidence scores in relation to participants' free-form comments. As we will show, participants in the treatment condition reported more cautious argumentation in their explanations, and that was linked to their decreased self-reported confidence. No significant effects were observed for *decision time*. Lastly, while we intentionally selected models of comparable quality, it was insightful to examine participants' *decision changes* with respect to their initial model choices after they interacted with RiskRAG versus the baseline. Developers changed their mind more with RiskRAG (82% vs. 76%), as well designers (70% vs. 62%), but not media professionals (73% vs. 78%).

Regarding preference between the reports, the RiskRAG report was favored by 58% of developers, 63% of designers, and 70% of media professionals.

**Table 6: Results for decision confidence from the final user study. A 2x2 analysis based on two factors: (baseline vs. treatment) and phase (pre-report vs. post-report). Significant results are highlighted in bold. Confidence dropped significantly from before to after seeing the risk report for developers and media professionals, with a bigger drop after seeing RiskRAG compared to the baseline. \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$**

Metric	Factors	AI Developers (n=38)	UX Designers (n=40)	Media Professionals (n=37)
Decision confidence	Condition (Main effect)	F(1,37) = 0.39	F(1,39) = 0.00	F(1,36) = 2.26
	<b>Phase (Main effect)</b>	<b>F(1,37) = 6.13*</b>	F(1,39) = 0.09	<b>F(1,36) = 5.10*</b>
	Condition x Stage (Interaction)	F(1,37) = 0.59	F(1,39) = 1.06	F(1,36) = 2.03



**Figure 7: Results for explanation quality ( $y$ -axis) from the final user study in three cohorts: developers, designers, and media professionals. Mean explanation quality ( $\mu$ ) improved for all cohorts after using RiskRAG, as opposed to the baseline reports, with statistically significant increases as shown by Wilcoxon test results ( $W, p$ ) for developers and designers. \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ . After using the RiskRAG report, participants identified higher number of relevant risks and mitigation strategies and provided better quality explanations for their model selection compared to baseline.**

*Qualitative Results.* Participant preferences for the RiskRAG report revealed five key themes. Below we describe these themes along with major sub-themes.<sup>26</sup>

**Supporting decision-making:** In this major theme, participants appreciated the report’s ability to simplify the cognitive load for analysis as it provided clear structure, prioritized risks, and revealed stakeholder impacts, enabling effective comparison across models and their decision-making. “Risk heatmaps are effective for summarizing complex data and can be especially useful in fast-paced decision-making scenarios.” (P0, dev).<sup>27</sup> “I can understand easily what is good and what is bad, to compare and make a decision easier.” (P11, des). A key sub-theme that emerged across all cohorts was **cautious decision-making**. Participants expressed greater awareness of potential risks, leading to more deliberate choices: “I’ll proceed

cautiously with the use of MODELX given the serious ethical and reputational risks highlighted by past incidents.” (P1, des) “I would deliberately undertake a training to the developers and the teams that will be involved in generating the content on information sensitivity, reputation and consistency of information.” (P4, dev) “This model, with its larger scale and instruction-based design, has the potentials to produce more accurate and nuanced description [...] However, even a large-scale model must be thoroughly tested for bias, especially in diverse contexts, to avoid making harmful associations.” (P35, med) Clear, contextualized risk reports emphasizing real-world harms led participants to approach model selection more carefully. Explicit examples of past incidents increased awareness of potential consequences, prompting them to reconsider choices initially based solely on technical details. This shift in awareness contributed to a notable reduction in decision confidence in the RiskRAG group during the post-report phase. Although both groups showed decreased confidence after reviewing risk reports, we did not see detailed

<sup>26</sup>The full codebook with all sub-themes is in [Supp. Mat. 5.3](#).

<sup>27</sup>We represent participants by their ID and cohort, dev: developers, des: designers, med: media professionals.

deliberation in the baseline condition. By providing tangible examples of harm, RiskRAG encouraged more thoughtful, self-reflective decision-making, shifting model selection from a purely technical focus to a more cautious, risk-aware approach.

**Effective risk communication:** Participants valued consistent, comprehensive, actionable insights about potential risks. *“It has a much more detailed and comprehensive analysis and also a balanced assessment.”* (P5, med). *“I choose Risk Heatmap because it is a valuable tool for identifying, assessing, and communicating risk visually and effectively with broad contexts.”* (P2, dev).

**Accessible presentation:** This theme emphasizes the role of visuals and accessible formatting in usability. *“I prefer risk report A [RiskRAG] because it is visually appealing, elaborate, and uses simple and understandable language. Risk Report B is shallow and plain, and does not give people the motivation to read and comprehend what it reports.”* (P28, med). Even developers, the most knowledgeable cohort, acknowledged that the report simplified technical jargon, making it easier to understand. *“The risks were clearly outlined... which helps demystify the technical jargon and allow users to make ethical decisions.”* (P30, dev).

**Risk prioritization for mitigations:** Participants noted that the structured categorization supported better prioritization and planning. *“I prefer risk report A because the heatmap type made it easier to spot key risks quickly and prioritize actions.”* (P36, des).

**Initial learning curve:** A challenge emerged, with some participants citing the initial cognitive load required to familiarize themselves with the report. *“Though understanding the heatmap was a bit of a struggle, it gets easier once you understand.”* (P1, dev).

For the users who preferred the baseline, two key themes emerged, highlighting potential areas for improvement for RiskRAG.

**Detailed textual explanations:** baseline allowed easier interpretation for some participants. *“The text-based format allows for more comprehensive explanations of the risks and mitigations.”* (P34, des). *“The structured text clarified the nuances and provided specific examples, making it easier to evaluate the ethical implications.”* (P22, dev).

**Familiarity and experience:** Participants highlighted the comfort they felt with the familiar textual style of baseline risk content, aligning with their prior experiences in similar tasks. *“Because it was something I had encountered before this study... made it easy for me to express my thoughts.”* (P11, dev).

## 5.4 Ethics

All three studies, including co-design and evaluation, were approved by the authors’ organization. Participants received informed consent detailing the study’s purpose, data usage, and their rights. Confidentiality and anonymity were ensured, with data handled according to established ethical guidelines. Participants recruited via Prolific were compensated at a minimum rate of \$12/hour. Research materials, such as study artifacts and thematic analysis codebook were shared in compliance with transparency criteria outlined by Salehzadeh Niksirat et al. [61].

## 6 Discussion

In this study, we introduced RiskRAG, a Retrieval Augmented Generation-based system designed to improve the risk reporting

process for AI models. Our work addresses a gap in current AI model documentation practices, particularly in model cards, which often lack comprehensive and specific risk assessments tailored to both the model and AI uses. Through iterative co-design and user studies with a total of 181 AI developers, UX designers, and media professionals, we empirically demonstrated that RiskRAG provides more contextualized and actionable risk reports, compared to the baseline model card risk sections. RiskRAG encouraged a more cautious and deliberative approach to model selection, effectively supporting decision-making.

### 6.1 Data-driven Solution for Risk Documentation for AI developers

In contrast to previous efforts that leverage LLMs to envision AI risks [11, 29, 69], RiskRAG is grounded in real-world datasets, ensuring a robust and adaptable solution. It employs retrieval-augmented generation to source human-written risks from model cards or documented real-world harms, minimizing hallucinated or generic risks. RiskRAG enhances existing solutions that replicate predefined formats or focus on subsets of model cards (e.g., CradGen tied to research papers or GitHub repositories [39]). Instead, it works on all model cards, addressing critical gaps in current methodologies for risk reporting as highlighted in Table 7. Risk Cards [17] do not include model-specific risks or mitigation strategies, Kennedy-Mayo and Gord [34] omit risk categorization, and neither documentation solutions address risk prioritization. Current solutions offer partial solutions but fail to meet the comprehensive requirements of risk documentation identified. ExploreGen [29] generates uses but not risks, while AHA! [11] and FarSight [69] produce risks for abstract AI systems without tailoring them to specific models or prioritizing them. Additionally, none propose actionable mitigation strategies. In contrast, RiskRAG uniquely prioritizes risks based on real-world harms, while associating each risk with tailored mitigation strategies. However, we believe these previous solutions can complement RiskRAG. For instance, RiskRAG did not show statistical significance in participant responses regarding whether risks were adequately contextualized to the use case or not. Incorporating tools like FarSight, which generates risks for specific applications, could enhance RiskRAG’s ability to provide a more comprehensive and context-sensitive risk assessment.

RiskRAG demonstrates strong potential for *generalizability*, offering two key benefits for unseen and lesser-known models: its structured template prompts developers to generate meaningful risk content, and the generated reports serve as a valuable starting point compared to a blank slate. To assess this capability, we selected four lesser-known models from the 450K snapshot of models from HuggingFace (selection process and the models described in Appendix E). Manual evaluation of their RiskRAG-generated reports confirmed the utility and relevance of the outputs. These generated reports, shared in [Supp. Mat. 3.2](#), illustrate how RiskRAG can support effective risk documentation even for models lacking extensive prior information, reinforcing its adaptability and broader applicability.

Further, RiskRAG’s architecture is designed for *scalability*, enabling seamless integration with larger and more comprehensive

**Table 7: Comparison of RiskRAG to closely related prior works in two main research areas: AI risk documentation, and tools for populating such documentation. For the AI risk documentation solutions, we examine whether they address each of our identified design requirements and whether they are specifically designed for model cards. For tools populating the AI risk documentation, we assess whether their content meets the requirements, is tailored to model cards, and whether it leverages RAG techniques. AI documentation proposals, Risk Cards and Model Cards in 2024 lack prioritization of risks (R5), do not address mitigation strategies (R4), or structure them according to taxonomies (R2). Tools for populating AI risk documentation such as ExploreGen, AHA!, FarSight, and CardGen do not focus on generating model-specific risks (R1), mitigation strategies (R4), or to prioritize them (R5), and they mainly rely on LLMs only. While CardGen uses RAG techniques, it produces risk content mimicking existing model cards, which itself falls short of the design requirements.**

	Model-specific risks (R1)	Structured risks (R2)	Contextualized to uses (R3)	Mitigation strategies (R4)	Prioritization of risks (R5)	Designed for model cards	Uses RAG
AI Risk Documentation							
Risk Cards [17]	✗	✓	✓	✗	✗	✗	-
Model Cards 2024 [34]	✓	✗	✓	✓	✗	✓	-
Tools for Populating AI Risk Documentation							
ExploreGen [29]	✗	✗	✓	✗	✗	✗	✗
AHA! [11]	✗	✓	✓	✗	✗	✗	✗
Farsight [69]	✗	✓	✓	✗	✓	✗	✗
CardGen [39]	✓	✗	✗	✗	✗	✓	✓
<b>RiskRAG (ours)</b>	✓	✓	✓	✓	✓	✓	✓

datasets, such as the MIT AI Risk Repository<sup>28</sup> [64], the Automation Incident Repository (AIAAIC) Repository<sup>29</sup>, and the OECD AI Incident Monitoring system (AIM)<sup>30</sup>. This ensures that RiskRAG remains flexible and future-proof, evolving as new data emerges. Additionally, RiskRAG reports offer clear and actionable mitigation strategies for each identified risk, empowering model users to address potential issues proactively before deploying them. Importantly, we envision RiskRAG not as a final solution but as a tool to support AI developers in producing effective risk reports. Starting with its generated content and structured format, developers can update, omit irrelevant risks, and be inspired to produce new ones, fostering a collaborative and iterative approach to AI risk documentation. This process not only streamlines risk assessment but also addresses the challenge of limited motivation for AI developers to identify potential harms [42, 43] by providing a structured, accessible foundation for risk documentation.

## 6.2 Raising Awareness and Promoting Responsible AI Use

As of July 2024, out of 450K model cards on HuggingFace, only 64K included risk-related sections, and just 2672 of those were unique. This means that approximately 86% of model cards on HuggingFace do not mention any risks. These numbers are consistent with findings from previous studies [4, 37]. Our co-design study supports these results, as AI developers reported that they typically focus on the technical aspects of model documentation, often overlooking risk-related sections. However, they also reported feeling enlightened and inspired by the risk content we provided during the study, in alignment with research showing that even AI practitioners and researchers find it challenging to anticipate the risks associated with AI systems and models [8, 19].

The practical potential of RiskRAG lies in its integration with platforms such as HuggingFace, GitHub, Model Zoo<sup>31</sup>, PyTorch Hub<sup>32</sup>, or Google AI Hub<sup>33</sup>. It can serve as a template of or an interactive interface for model card creation, enabling AI developers to prioritize model-specific risks and mitigations contextualized to diverse uses. Through an interactive process, developers can refine risk assessments, retrieve relevant examples of AI incidents, and identify mitigation strategies with reduced effort. Crowdsourced feedback could further enhance RiskRAG, refining its prioritization techniques and producing tailored, actionable reports over time.

Our findings highlight RiskRAG’s ability to foster deliberative and cautious decision-making. The preliminary study confirmed that RiskRAG met all desired requirements, while the final study demonstrated its effectiveness in enhancing users’ ability, developers, designers, and media professionals alike, to identify risks, devise mitigations, and improve explanation quality. By deepening users’ understanding of model impacts, RiskRAG can enable more informed decision-making, such as opting against unsuitable models, strengthening risk management, and making critical adaptations prior to deployment. This approach helps prevent harmful incidents and promotes ethical AI use, ensuring AI technologies are developed and deployed in alignment with responsible and safe practices.

## 6.3 Broader Implications: Model Cards, Public Outreach, and Policy Making

Recent efforts have sought to enhance the ethical considerations and risk sections in model cards [17, 34]. The current standard for documenting AI models—model cards—can be significantly enhanced through the integration of insights from our work with RiskRAG. Specifically, the risk sections could be transformed to

<sup>28</sup><https://airisk.mit.edu/>

<sup>29</sup><https://www.aiaaic.org/aiaaic-repository>

<sup>30</sup><https://oecd.ai/en/incidents>

<sup>31</sup><https://modelzoo.co/>

<sup>32</sup><https://pytorch.org/hub/>

<sup>33</sup><https://cloud.google.com/vertex-ai/>

reflect the detailed risk assessments generated by RiskRAG, making risk documentation more comprehensive. This enhancement has the potential to establish a new best practice, encouraging deeper engagement with potential risks. Consequently, this could position model cards as robust foundations for impact assessment reports [5, 6, 44], enabling thorough evaluations before deployment. RiskRAG also opens new research avenues regarding the presentation of use-specific risks and mitigations in model documentation. By tailoring risk information to specific use cases, model cards could evolve to resemble impact assessment reports, offering a more structured approach to decision-making about the consequences of deploying AI models in various contexts.

Beyond risk sections, our user-centred approach to developing RiskRAG could inspire improvements across other sections of model cards, such as model and data specifications. By optimizing the presentation and usability of model cards, they can become more accessible, informative, and effective for developers at all levels.

As regulatory scrutiny [14, 30] around AI technologies increases, businesses and developers must ensure compliance with evolving laws. RiskRAG can streamline this process by identifying risks and aligning its reporting with specific regulatory requirements, reducing the complexity of navigating legal frameworks and helping organizations address potential compliance gaps proactively. This alignment ensures models meet regulatory standards from the outset. Moreover, RiskRAG has the potential to enhance the accessibility of risk information for both the *general public* and *polymakers*. Participants in our study highlighted concerns about the lack of transparency in traditional risk reports, perceiving missing or unclear information as intentional, affecting trust and accountability. As one participant noted, *“It seems in their documentation to pretend like they’re conveying actual information.”* (P25, dev), while another stated, *“It made me feel as though it was actually hiding information about its risks that it would rather have people not know.”* (P24, dev). By presenting clearer and more comprehensive insights, RiskRAG could foster greater trust, enabling stakeholders to make *better-informed* decisions and promoting more responsible AI governance.

## 6.4 Limitations

**RAG Model Biases and Hallucinations.** Retrieval-Augmented Generation (RAG), while better than Large Language Models (LLMs) at grounding generated information in external databases and reducing hallucinations, has limitations. Retrieval bias may occur if certain model cards are over-represented, though our user studies with diverse model types (e.g., text generation, image and text-to-text, speech-to-text) showed high-quality risk generation. By incorporating real-world incident risks, we further reduced reliance on model cards and captured undocumented risks. Frequency bias in risk prioritization, based on incident frequency, could lead to an over-focus on certain risks and under-focus on others, if the incident repository, in this case AIID, is biased. This can be mitigated by integrating multiple repositories such as the OECD AIM [54] and the AIAIC. Despite fewer hallucinations than LLMs [23, 55], RAG systems still occasionally generate incorrect risks. For example, RiskRAG incorrectly flagged “generates disinformation by creating misleading or false images” for a model that processes but does not

generate images (i.e., an image-to-text model). In our experiments, emphasizing model type, as well as types of input and output of the model in prompts, minimized such errors. Since RiskRAG is designed to assist, not replace, developers, these errors pose limited risk, as developers can and are expected to refine outputs. Future work could add a secondary generative agent to verify risks and further reduce hallucinations.

Lastly, we acknowledge that some risks extracted by our method may not be entirely relevant to a specific model. In our evaluation detailed in §5.1, we automatically validated the quality and relevance of the retrieved risks; however, due to the challenge of recruiting highly knowledgeable AI and ethical experts for specific models, we could not conduct a large-scale expert study on this issue. Instead, we manually assessed the quality of risks for two sets of models: four models for which the three authors had high expertise, and four less-known models (discussed in Appendix E). When asked, *“Are the risks relevant to the model and its type?”* on a scale from 1 to 5, the average response across all models was 4.75, indicating strong agreement on their relevance.

**Limitations of Evaluation Data.** We selected the most popular model cards for our pseudo ground truth dataset, which may inadvertently favor newer models that are closely aligned with existing and well-known models. To verify RiskRAG’s performance on lesser-known models, we conducted an additional experiment (Appendix E), which confirmed its effectiveness. Future work could explore incorporating additional information specific to the new models (e.g., its associated paper, similar to CardGen) to enhance output quality in such cases.

**Artificial Setting and Study Scope.** Although we worked with 165 AI stakeholders across two studies and provided realistic tasks, the studies were conducted in controlled settings rather than real-world environments, and they were one-time experiments rather than longitudinal. This limits the external validity of our findings, especially concerning RiskRAG’s long-term performance.

**Learning Curve and Adapting to RiskRAG.** Thematic analysis revealed that some participants favoring RiskRAG initially faced challenges in adapting to it, while those who preferred baseline reports often cited familiarity as the key factor for their choice. This disparity, alongside preference scores from the final study—58% for developers compared to 70% for media professionals—suggests that AI stakeholders accustomed to traditional reports may require additional effort to adjust to RiskRAG. Future research should explore a hybrid approach, blending RiskRAG’s visually structured matrix with textual explanations and tabular examples common in existing reports, as proposed by some of our participants: *“Although the heatmap allows you to see at a glance, I think I prefer the text... I think the ideal would be a combination of both”* (P6, des).

**Incomplete (Systemic) Risk Coverage.** Although participants perceived RiskRAG’s reports to be more detailed, there is a chance that risks derived from similar models could still be incorrect, as discussed above. Furthermore, systemic risks, which take longer to materialize, are difficult to fully capture even with real-world incident data [68]. Thus, RiskRAG’s coverage of systemic harms may still be incomplete.

**Pretrained Models.** We used pretrained embedding models for RiskRAG without further fine-tuning, based on prior work suggesting this approach generally works well. However, given the complexity of generating detailed risk content, future research should explore whether *pretraining RAG components* on model card data or real-world incidents could improve the quality of risk assessments.

**Static Presentation.** We focused on static PDF presentations of model cards, but it is likely that an interactive format would be even more effective. Future work should explore how developers interact with dynamic, interactive versions of RiskRAG, where risks and mitigations adapt based on the selected use case. Additionally, a community-driven feedback mechanism could invite developers to report new risks encountered during production, enriching the risk database over time.

## 7 Conclusion

Our work on RiskRAG demonstrates the potential of a data-driven, AI-assisted risk reporting system that aligns with the needs of AI developers. By addressing gaps in current model card risk reporting and providing actionable insights, RiskRAG fosters responsible AI use and improves risk documentation practices. With further refinement and adoption, RiskRAG could significantly enhance how AI models are evaluated and deployed in the real world, contributing to safer and more transparent AI systems.

## References

- [1] Ada Lovelace Institute. 2022. *Algorithmic Impact Assessment: AIA Template*. Ada Lovelace Institute. Retrieved January 22, 2024 from <https://www.adalovelaceinstitute.org/resource/aia-template/>
- [2] Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual Embeddings: When Are They Worth It? <https://doi.org/10.48550/arXiv.2005.09117> arXiv:2005.09117 [cs]
- [3] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.
- [4] Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L.C. Guo. 2023. Aspirations and Practice of ML Model Documentation: Moving the Needle with Nudging and Traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. <https://doi.org/10.1145/3544548.3581518>
- [5] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. AI Design: A Responsible Artificial Intelligence Framework for Prefilling Impact Assessment Reports. *IEEE Internet Computing* 28, 5 (Sept. 2024), 37–45. <https://doi.org/10.1109/MIC.2024.3451351>
- [6] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-Designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 168–180. <https://doi.org/10.1609/ai.es.v7i1.31627>
- [7] Edyta Bogucka, Sanja Šćepanović, and Daniele Quercia. 2024. Atlas of AI Risks: Enhancing Public Understanding of AI Risks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 12 (Oct. 2024), 33–43. <https://doi.org/10.1609/hcomp.v12i1.31598>
- [8] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. <https://doi.org/10.48550/arXiv.2011.13416> arXiv:2011.13416 [cs]
- [9] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- [10] Virginia Braun and Victoria Clarke. 2012. Thematic Analysis. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*. American Psychological Association, Washington, DC, US, 57–71. <https://doi.org/10.1037/13620-004>
- [11] Zana Bućinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. <https://doi.org/10.48550/arXiv.2306.03280> arXiv:2306.03280 [cs]
- [12] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS - Meta AI Literacy Scale: Development and Testing of an AI Literacy Questionnaire Based on Well-Founded Competency Models and Psychological Change- and Meta-Competencies. *Computers in Human Behavior: Artificial Humans* 1, 2 (Aug. 2023), 100014. <https://doi.org/10.1016/j.chbah.2023.100014>
- [13] Jiyoo Chang and Christine Custis. 2022. Understanding Implementation Challenges in Machine Learning Documentation. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–8. <https://doi.org/10.1145/3551624.3555301>
- [14] Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 25 July 2024. *Official Journal of the European Union* L 168 (2024), 1–15. [https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L\\_202401689](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689) Accessed: 2024-07-29.
- [15] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FACT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. <https://doi.org/10.1145/3531146.3533108>
- [16] Wesley Hanwen Deng, Solon Barocas, and Jennifer Wortman Vaughan. 2025. Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing the Potential Negative Societal Impact of Their Work. <https://doi.org/10.1145/3711076> arXiv:2408.01057 [cs]
- [17] Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. 2023. Assessing Language Model Deployment with Risk Cards. arXiv:2303.18190 [cs]
- [18] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. 2023. Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. *Information Fusion* 99 (Nov. 2023), 101896. <https://doi.org/10.1016/j.inffus.2023.101896>
- [19] Kimberly Do, Rock Yuren Pang, Jiachen Jiang, and Katharina Reinecke. 2023. “That’s Important, but...”: How Computer Science Researchers Anticipate Unintended Consequences of Their Research Innovations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3544548.3581347>
- [20] OpenAI et al. 2024. GPT-4 Technical Report. <https://doi.org/10.48550/arXiv.2303.08774> arXiv:2303.08774 [cs]
- [21] Sabri Eyuboglu, Karan Goel, Arjun Desai, Lingjiao Chen, Mathew Monfort, Chris Ré, and James Zou. 2024. Model ChangeLists: Characterizing Updates to ML Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2432–2453. <https://doi.org/10.1145/3630106.3659047>
- [22] Lynn Frewer. 2004. The Public and Effective Risk Communication. *Toxicology Letters* 149, 1-3 (April 2004), 391–397. <https://doi.org/10.1016/j.toxlet.2003.12.049>
- [23] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs]
- [24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
- [25] Delaram Golpayegani, Isabelle Hupont, Cecilia Panigutti, Harshvardhan J. Pandit, Sven Schade, Declan O’Sullivan, and Dave Lewis. 2024. AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. In *Privacy Technologies and Policy: 12th Annual Privacy Forum, APF 2024, Karlstad, Sweden, September 4–5, 2024, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 48–72. [https://doi.org/10.1007/978-3-031-68024-3\\_3](https://doi.org/10.1007/978-3-031-68024-3_3)
- [26] Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. 2023. To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act’s High-Risk AI Applications and Harmonised Standards. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 905–915. <https://doi.org/10.1145/3593013.3594050>
- [27] Google. 2024. Model cards: A step towards AI transparency. <https://modelcards.withgoogle.com/about>. Accessed: 2024-07-29.
- [28] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* 18, 1 (2006), 59–82. <https://doi.org/10.1177/1525822X05279903>
- [29] Viviane Herdel, Sanja Šćepanović, Edyta Bogucka, and Daniele Quercia. 2024. ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies. <https://doi.org/10.48550/arXiv.2407.12454> arXiv:2407.12454 [cs]
- [30] The White House. 2022. Blueprint for an AI Bill of Rights | OSTP. <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>.
- [31] Isabelle Hupont, David Fernández-Llorca, Sandra Baldassarri, and Emilia Gómez. 2024. Use Case Cards: A Use Case Reporting Framework Inspired by the European AI Act. *Ethics and Information Technology* 26, 2 (March 2024), 19. <https://doi.org/10.1007/s10676-024-09757-7>

- [32] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 560–575. <https://doi.org/10.1145/3442188.3445918>
- [33] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Hawai'i, 15696–15707.
- [34] DeBrae Kennedy-Mayo and Jake Gord. 2024. "Model Cards for Model Reporting" in 2024: Reclassifying Category of Ethical Considerations in Terms of Trustworthiness and Risk Management. <https://doi.org/10.48550/arXiv.2403.15394> arXiv:2403.15394 [cs]
- [35] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. 2002. AIMQ: a methodology for information quality assessment. *Information & management* 40, 2 (2002), 133–146.
- [36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., Vancouver, Canada, 9459–9474.
- [37] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. What's Documented in AI? Systematic Analysis of 32K AI Model Cards. arXiv:2402.05160 [cs]
- [38] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.
- [39] Jiarui Liu, Wenkai Li, Zhijiang Jin, and Mona Diab. 2024. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. arXiv:2405.06258 [cs]
- [40] LMSYS. 2024. *LMSYS Chatbot Arena Leaderboard*. Huggingface. Retrieved July 9, 2024 from <https://lmsys.org/blog/2023-06-22-leaderboard/>
- [41] Bill MacCartney and Christopher D. Manning. 2008. Modeling Semantic Containment and Exclusion in Natural Language Inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Donia Scott and Hans Uszkoreit (Eds.). Coling 2008 Organizing Committee, Manchester, UK, 521–528.
- [42] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 52:1–52:26. <https://doi.org/10.1145/3512899>
- [43] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376445>
- [44] Alessandro Mantelero and Maria Samantha Esposito. 2021. An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review* 41 (2021), 105561.
- [45] Sean McGregor. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 15458–15463. <https://doi.org/10.1609/aaai.v35i17.17817>
- [46] Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: A Survey. arXiv:2302.07842 [cs]
- [47] Matthew B. Miles, A. Michael Huberman, and Johnny Saldana. 2013. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, CA, USA.
- [48] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 220–229. <https://doi.org/10.1145/3287560.3287596>
- [49] Saif M Mohammad. 2022. Ethics sheet for AI tasks. *Computational Linguistics* 48, 2 (2022), 239–278.
- [50] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2014–2037. <https://doi.org/10.18653/v1/2023.eacl-main.148>
- [51] National Institute of Standards and Technology (NIST). 2023. AI Risk Management Framework. <https://www.nist.gov/itl/ai-risk-management-framework> Accessed: 2024-11-19.
- [52] José Luiz Nunes, Gabriel D. J. Barbosa, Clarisse Sieckenius de Souza, Helio Lopes, and Simone D. J. Barbosa. 2022. Using Model Cards for Ethical Reflection: A Qualitative Exploration. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems (IHC '22)*. Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3554364.3559117>
- [53] OECD. 2020. Guidelines for MNEs - Organisation for Economic Co-operation and Development. <https://mneguidelines.oecd.org/rbc-and-digitalisation.htm>.
- [54] OECD. 2024. AI Incidents. <https://oecd.ai/en/incidents> Accessed: 2024-11-27.
- [55] Oded Ovadia, Menachem Brief, Moshik Misha'el, and Oren Elisha. 2024. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. <https://doi.org/10.48550/arXiv.2312.05934> arXiv:2312.05934 [cs]
- [56] David Piorowski, Michael Hind, and John Richards. 2024. Quantitative AI Risk Assessments: Opportunities and Challenges. <https://doi.org/10.48550/arXiv.2209.06317> arXiv:2209.06317 [cs]
- [57] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 913–926. <https://doi.org/10.1145/3600211.3604712>
- [58] Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. 2024. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 1200–1217. <https://doi.org/10.1609/aies.v7i1.31717>
- [59] Athena Richards, Sheila Convery, Margaret O'Mahony, and Brian Caulfield. 2024. Pre and post Covid preferences for working from home. *Travel Behaviour and Society* 34 (2024), 100679.
- [60] Johnny Saldana. 2015. *The Coding Manual for Qualitative Researchers*. SAGE, CA, USA.
- [61] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanatham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–23. <https://doi.org/10.1145/3544548.3580848>
- [62] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. <https://doi.org/10.48550/arXiv.1910.01108> arXiv:1910.01108 [cs]
- [63] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 440–451. <https://doi.org/10.1145/3531146.3533110>
- [64] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. <https://doi.org/10.48550/arXiv.2408.12622> arXiv:2408.12622 [cs]
- [65] Bernd Carsten Stahl, Josephina Antoniou, Nitika Bhalla, Laurence Brooks, Philip Jansen, Blerta Lindqvist, Alexey Kirichenko, Samuel Marchal, Rowena Rodrigues, Nicole Santiago, Zuzanna Warso, and David Wright. 2023. A Systematic Review of Artificial Intelligence Impact Assessments. *Artificial Intelligence Review* 56, 11 (2023), 12799–12831. <https://doi.org/10.1007/s10462-023-10420-8>
- [66] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. <https://doi.org/10.48550/arXiv.2104.08663> arXiv:2104.08663 [cs]
- [67] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. 2021. ECCOLA – A Method for Implementing Ethically Aligned AI Systems. *Journal of Systems and Software* 182 (Dec. 2021), 111067. <https://doi.org/10.1016/j.jss.2021.111067>
- [68] Julia De Miguel Velázquez, Sanja Šćepanović, Andrés Gvrtiz, and Daniele Quercia. 2024. Decoding Real-World Artificial Intelligence Incidents. *Computer* 57, 11 (Nov. 2024), 71–81. <https://doi.org/10.1109/MC.2024.3432492>
- [69] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–40. <https://doi.org/10.1145/3613904.3642335>
- [70] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *J. Mach. Learn. Res.* 24, 1 (Jan. 2023), 257:12058–257:12065.
- [71] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986 [cs]

- [72] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. <https://doi.org/10.18653/v1/N18-1101>
- [73] Xinyu Yang, Weixin Liang, and James Zou. 2024. Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face. <https://doi.org/10.48550/arXiv.2401.13822> arXiv:2401.13822 [cs]
- [74] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., Vancouver, Canada, 27263–27277.
- [75] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. <https://doi.org/10.48550/arXiv.1904.09675> arXiv:1904.09675
- [76] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lema Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models. <https://doi.org/10.48550/arXiv.2309.01219> arXiv:2309.01219 [cs]

## A List of papers selected from literature to elicit design requirements for risk reporting

- (1) *Model cards for model reporting* [48]. The paper introduces the concept of model cards, aimed at providing transparent documentation of AI models. The ethical considerations section was intended to demonstrate the ethical considerations that went into model development, surfacing ethical challenges and risks, and the mitigation strategies that were used. Model cards should identify potential risks and harms associated with the model’s usage. It should explicitly state the primary intended uses of the AI model. This helps users understand the scope and limitations of the model, reducing the risk of misuse.
- (2) *Interactive model cards: a human-centered approach to model documentation* [15]. This paper discusses enhancing model documentation, specifically through interactive model cards. They find that current risk sections are ambiguous and the topics of safety and ethics were too abstract.
- (3) *Using model cards for ethical reflection: a qualitative exploration* [52]. This paper discusses the role of model cards in ethical reflection, which is crucial for understanding and documenting AI risks. The paper finds that developers selectively document ethical concerns in AI model cards, highlighting potential risks of incomplete ethical reflection in AI development. This suggests the need for better documentation practices to ensure more ethically informed AI design.
- (4) *Aspirations and practice of ML model documentation: Moving the needle with nudging and traceability* [4]. This paper focuses on the gaps between proposed model documentation practices and actual practices. They found that only about 35% of models’ documentation had a discussion about bias or ethics and only 10% about mitigating them.
- (5) *Understanding implementation challenges in machine learning documentation* [13]. This paper addresses the challenges in implementing ML documentation, which is essential for understanding the hurdles in reporting AI risks. They suggested making documentation a project deliverable to incentivize better practices.
- (6) *Model ChangeLists: Characterizing updates to ML models* [21]. This paper explores documenting updates to ML models, which relates directly to maintaining and reporting AI risks throughout the model lifecycle.
- (7) *What’s documented in AI? Systematic Analysis of 32K AI Model Cards* [37] An analysis of 32K model cards from HuggingFace revealed that only 17% of all cards and 39% of the top 100 most downloaded included sections on risks and limitations. They found that model cards report the limitations of the data used for training and the technical and societal limitations of the AI model.
- (8) *"Model Cards for Model Reporting" in 2024: Reclassifying Category of Ethical Considerations in Terms of Trustworthiness and Risk Management* [34]. proposed restructuring the ethical considerations section to clearly outline regulatory, reputational, and operational risks.

## B Reporting artifacts generated during the co-design process

We provided an overview of the iterative development of the risk report during our co-design process in Figure 8.

## C RiskRAG and CardGen

We compared RiskRAG with CardGen using the CardBench evaluation set (Table 8). Of the 294 model cards in CardBench, only 40 contain any risk-related content (referred to as “bias” in [39]). For a fair evaluation, we focused on these 40 model cards. To evaluate RiskRAG, we adapted the ground-truth model cards in CardBench to focus on individual risks, as detailed in §5.1.2 and illustrated in Figure 5. For CardGen, we assessed its performance solely on the risk-related content (“bias”) in CardBench, consistent with the evaluation reported in [39]. For both approaches (RiskRAG and CardGen), we reported the same automated metrics applied to CardGen as in [39]: ROUGE [38], BERTScore [75], BARTScore [74], and NLI-finetuned models [41, 72].

As shown in Table 8, RiskRAG consistently matches or outperforms CardGen across all metrics. Although the formats of the risk-related content vary between the two approaches, we believe the evaluation results remain informative, demonstrating that RiskRAG provides competitive risk-related content.

## D Qualitative analysis of feedback from Preliminary Study

Two authors conducted an inductive thematic analysis [9] of qualitative feedback from open-ended questions using established qualitative coding methodologies [47, 60]. Participant responses were documented as sticky notes, and themes were collaboratively developed based on this data. The authors resolved disagreements through discussion, ensuring consensus. Each identified theme was supported by quotes from at least two participants, demonstrating data saturation [28].

Having established that the participants preferred RiskRAG report to the baseline one, we thematically analyzed the content of their qualitative responses to learn why. In their answers to the preference explanation, participants praised the RiskRAG report for (R1) comprehensive detail and depth of information (P11: *‘It is also a lot more detailed, saving me the time to answer follow-ups potentially.’*); (R2) clarity and structure (P28: *‘The tabular view made it much easier to understand the level of risk per use case and to quickly see if my use case was a high risk.’*); (R3) context-specific relevance (P3: *‘I preferred Risk report A because the risks were presented and categorized depending on whether they were applicable to the use cases being shown or not.’*); (R4) actionable mitigation strategies (P36: *‘Report B was much more helpful because it specifically spelled out the potential ethical and safety hazards and potential solutions for tackling them.’*); and (R5) prioritization of risk information (P2: *‘It more clearly prioritized certain risks and their real-world harm so that more important risks could be more focused on.’*). In addition to these themes relating our design requirements, also the following three themes emerged in participant responses: (1) usability for decision-making and communication (P7: *‘There is plenty of information stated in A to make an informed and helpful email.’* and P26: *‘Risk report A helps with the task more.’*); (2) visuals and layout

(P19: *‘The one with the graphics is better because it is more visual and allows you to consume much more information immediately.’* and P29: *‘Risk report A provides a more visual overview of factors that is easier to scan and comprehend.’*); and (3) trust in transparency compared to the baseline report (P24: *‘It made me feel as though Risk report B was actually hiding information about its risks that it would rather have people not know.’*).

## E Generalizability of RiskRAG to unseen and lesser known models

To evaluate the generalizability of RiskRAG, we selected four lesser-known models from the snapshot of 461,181 model cards on HuggingFace. Using cosine similarity, we compared the names of these models against 2.6K model names in our dataset (4.1.1), quantifying their semantic alignment. The four models with the lowest similarity scores were identified, ensuring they were among the least similar to those used to develop RiskRAG. This approach provided a diverse and challenging subset for testing RiskRAG’s robustness on novel and less familiar cases. The selected models<sup>34</sup> spanned various types—text-to-image, text-classification and text-generation—and had low usage, with downloads ranging from 5 to 60. The generated risk reports are available on [Supp. Mat. 3.2](#).

None of the four selected models originally included any risk-related sections. In contrast, RiskRAG generated 21, 25, 21, and 19 risks for each model, respectively. We manually inspected the reports of these models and found them not only relevant, but providing substantial support for creating otherwise missing risk sections. For instance, the text-to-image model CyberHarem/toyokawa\_fuuka\_theidolmstermillionlive, which generates NSFW anime characters, received a report with 25 relevant risks—13 of which were sourced from the AIID dataset. Identified risks included: “produces illegal content due to inclusion of CSAM in the dataset”, “promotes abusive violent or pornographic materials if misused”, “reinforces or exacerbates social biases” and “damages reputations by associating individuals or groups with offensive content”. This confirmed that the risks and mitigations in these reports were relevant, demonstrating RiskRAG’s ability to produce meaningful outputs even for lesser-known models.

To assess whether the identified design requirements were met for these lesser-known models, three authors independently rated the reports using the same evaluation criteria from the preliminary study (§5.2.3). The average scores across the five requirements were 4.17 (R1), 4.33 (R2), 4.00 (R3), 2.67 (R4), and 3.17 (R5). Most requirements were adequately satisfied, with scores similar to those in the user study (Figure 6). The lower score for mitigation strategies (R4) likely reflects a common pattern in model cards, where mitigation details are generally less thorough than the identified risks; and this gap is even more pronounced for lesser-known models, where both risks and mitigation strategies are often scarce or entirely absent.

<sup>34</sup><https://huggingface.co/artificialguybr/studioghlibi-redmond-2-1v-studioghlibi-lora-for-freedom-redmond-sd-2-1>, [https://huggingface.co/CyberHarem/toyokawa\\_fuuka\\_theidolmstermillionlive](https://huggingface.co/CyberHarem/toyokawa_fuuka_theidolmstermillionlive), <https://huggingface.co/MouezYazidi/XML-RoBERTa-CampingReviewsSentiment>, [https://huggingface.co/TahaCakir/enhanced\\_turkishReviews-generativeAI](https://huggingface.co/TahaCakir/enhanced_turkishReviews-generativeAI)

**Table 8: Evaluation of RiskRAG using ROUGE-L (R), BERTScore (BE), BARTScore (BA), and NLI pre-trained scorer (NLI). We provide CardGen scores on risk-related sections (termed ‘bias’ in [39]) for top-k = 10 as those are the only reported in [39] for comparison. RiskRAG matches or outperforms CardGen across all metrics, indicating RiskRAG could provide competitive risk-related content.**

Model ↓ Metric →	top-k = 5				top-k = 10				top-k = 15			
	R	BE	BA	NLI	R	BE	BA	NLI	R	BE	BA	NLI
Linq-Embed-Mistral	0.38	0.79	-3.30	0.69	0.36	0.75	-3.89	0.69	0.31	0.72	-4.15	0.65
SFR-Embedding-2_R	0.40	0.79	-3.31	0.69	0.37	0.76	-3.93	0.73	0.30	0.72	-4.28	0.73
bge-large-en-v1.5	0.37	0.78	-3.44	0.75	0.31	0.76	-3.84	0.70	0.22	0.71	-4.52	0.68
CardGen	-	-	-	-	0.20	0.59	-3.76	0.62	-	-	-	-

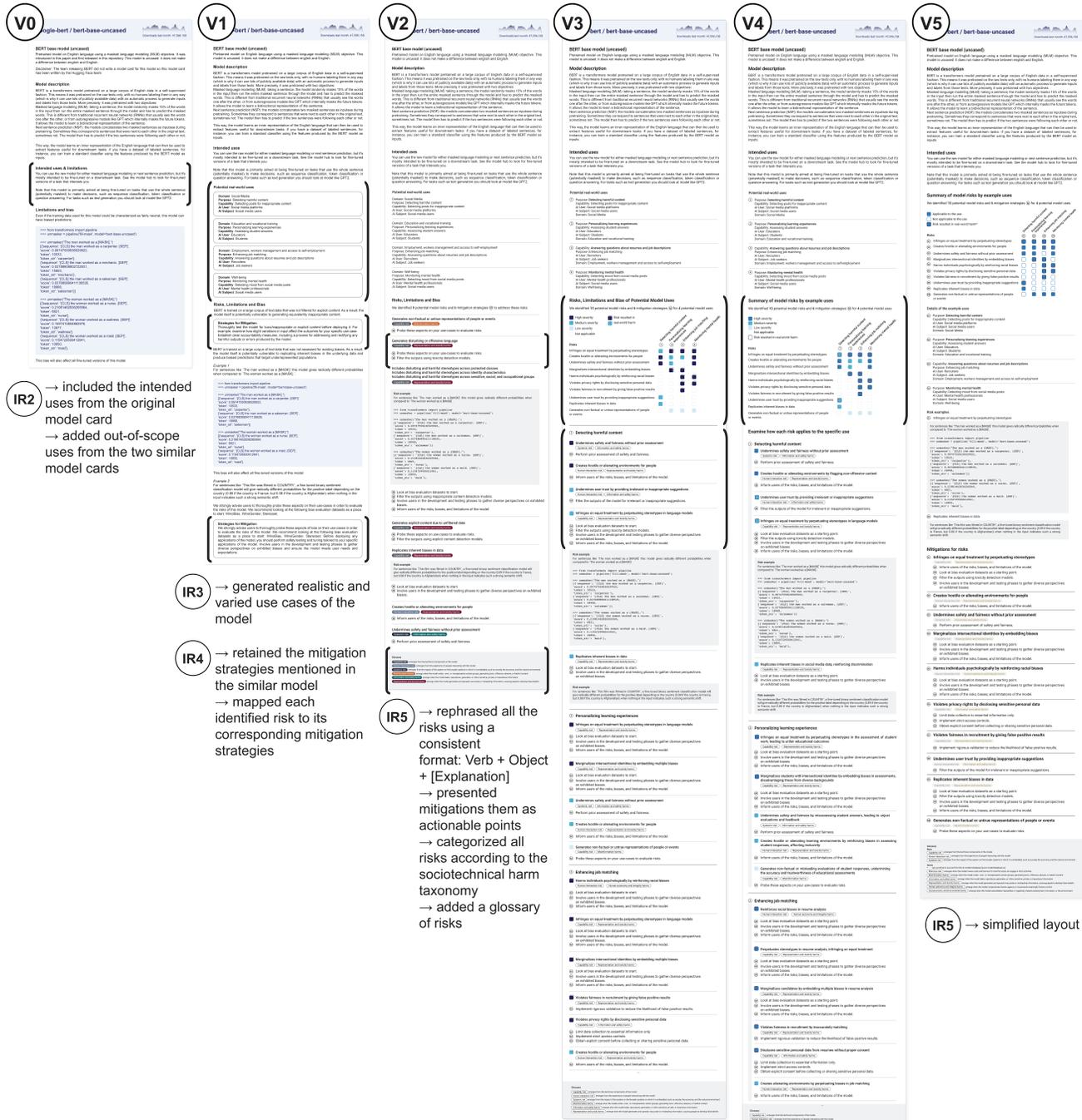
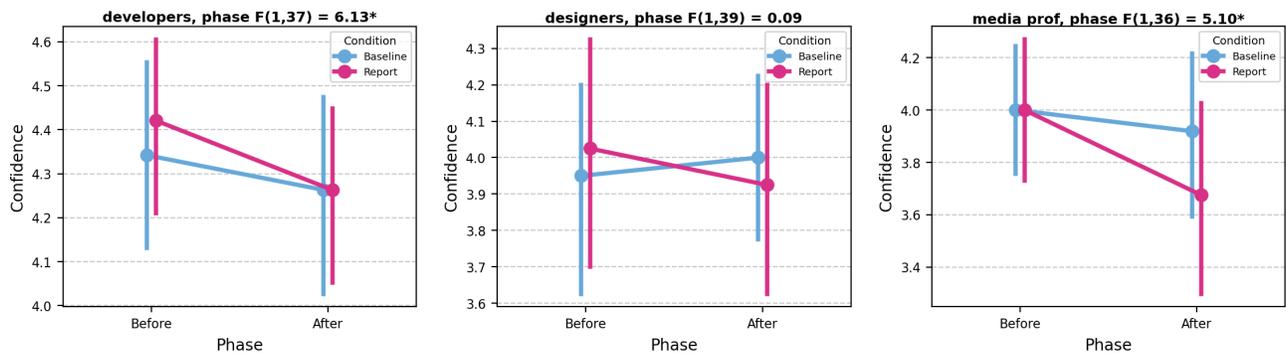


Figure 8: Overview of the iterative development of the risk report, highlighting changes in structure, presentation, and prioritization across five rounds.



**Figure 9: Final study: Differences in decision confidence (y-axis) before and after interacting with RiskRAG reports compared to the baseline reports for each of the three cohorts. The phase (before and after risk report) had a significant main effect on decision confidence for developers and media professionals, with confidence decreasing more for the RiskRAG report than the baseline ( $*p < 0.05$ ).**