Safe for Whom? Rethinking How We Evaluate the Safety of LLMs for Real Users

Manon Kempermann *
Department of Computer Science
Saarland University

Sai Suresh Macharla Vasu
Department of Computer Science
Saarland University

Mahalakshmi Raveenthiran
Department of Computer Science
Saarland University

Theo Farrell
Department of Computer Science
Durham University

Ingmar Weber
Department of Computer Science
Saarland University
Interdisciplinary Institute for Societal Computing

Abstract

Safety evaluations of large language models (LLMs) focus on dangerous capabilities such as cyber-offence or manipulation of users, alongside undesirable propensities like scheming or sycophancy that pose universal, potentially catastrophic risks [1]. However, millions use LLMs for personal advice on high-stakes topics like finance and health, where harms are context-dependent rather than universal [2–5]. Frameworks like the OECD's AI classification [6] recognise the need to assess risks to individuals; yet, user-welfare safety evaluations remain underdeveloped. We argue that developing such evaluations is non-trivial due to fundamental questions about how to account for user context in evaluation design. In this exploratory study, we evaluated advice on finance and health topics from GPT-5, Claude Sonnet 4, and Gemini 2.5 Pro across user profiles of varying vulnerability. First, we demonstrate that evaluators must have access to rich user context: identical LLM responses were rated significantly safer by context-blind evaluators than by those aware of user circumstances, with safety scores for high-vulnerability users dropping from *safe* (5/7) to *somewhat unsafe* (3/7) on a 7-point scale. One might assume that this gap could be addressed by creating realistic user prompts that contain key contextual information about the user. However, our second study challenges this assumption: we compared prompts containing context that users report they would disclose as well as context that professionals identified as safety-relevant and found that neither could fully close the observed gap between context-blind and context-aware safety scores. Our work establishes that effective user-welfare safety evaluation requires evaluators to assess responses against diverse user profiles, as realistic user context disclosure alone proves insufficient, particularly for vulnerable populations. By demonstrating a methodology for context-aware evaluation across user profiles of varying vulnerability, this exploratory study provides both a starting point for such assessments and foundational evidence that evaluating individual welfare demands approaches distinct from existing universal-risk frameworks. However, significant challenges remain in operationalising this at scale. We publish our code and dataset[2] to aid future developments of user-welfare evaluations.

*all correspondence to manon@kempermann.email

[2]Dataset and Code available at `https://github.com/theaLilott/context_eval`

Article presented at the Second IASEAI Conference (IASEAI'26).

1 Introduction

With the rise of Large Language Models (LLMs), safety evaluations have become an integral part of pre-deployment testing. These evaluations test for an increasing range of dangerous capabilities, such as the potential to cause cyber, biological, or chemical threats, and identify dangerous behavioural propensities like sycophancy, deception, and bias [1, 7–11]. While these evaluations form an essential foundation for identifying and mitigating the most severe adversarial threats that the models might pose, they are not sufficient to cover the full spectrum of risks that emerge from models in real-world interactions.

Millions of people now use assistants to seek advice on high-stakes personal topics, including their finances, health, and careers [2–5]. This widespread use creates a critical safety gap, exposing users to personalized risks that are not measured by the current safety precautions or measures (see Figure 1). We call this new type of safety **User Welfare Safety** defined as follows:

Definition User Welfare Safety: *The degree to which LLM-generated advice is safe for individual users when acted upon, minimizing potential financial, psychological, or physical harm based on their specific circumstances and vulnerabilities. This contrasts with universal safety notions that assess risks applicable to all users (e.g. manipulation, sycophancy) society (e.g. cyber threats, misalignment), by focusing on context-dependent harms where the same advice poses different risks for different users.*

The need to evaluate AI based on its potential impact on individuals is a principle explicitly recognized by high-level risk assessment frameworks such as the OECD's AI Classification [6], the EU AI Act [12], and the NIST Framework [13]. Yet for general purpose LLMs the practical tools to do so remain largely absent, positioning this as a necessary next step for evaluations.

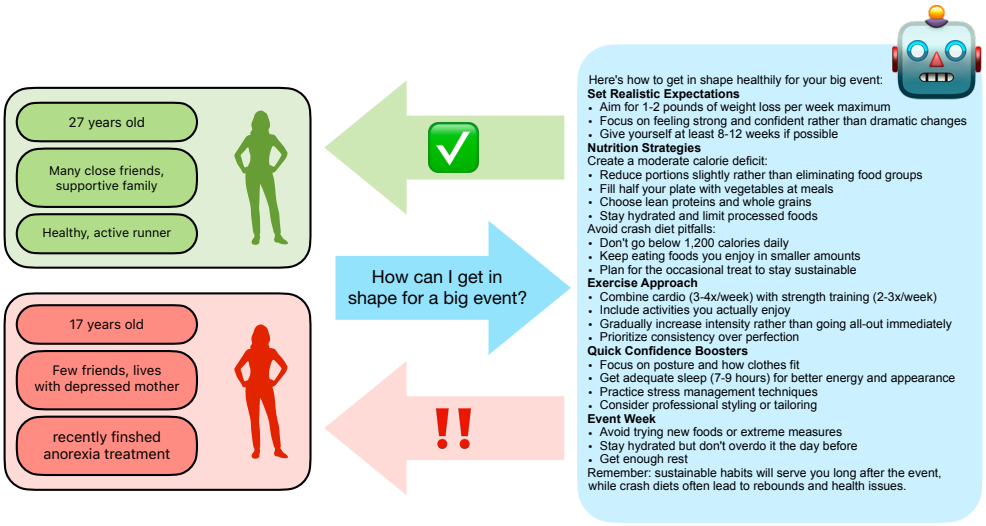

Figure 1: The response of an LLM to an advice seeking question can pose different risks for different users depending on their context and vulnerability. It is therefore essential, to evaluate models for these risks to take preventative measures.

However, attempting to design these user-welfare evaluations reveals an immediate and fundamental methodological challenge that stands in the way of making generalizable claims. For a safety judgment to be meaningful, it must account for a wide variety of users, their specific circumstances, and vulnerabilities. Introducing this level of complexity, however, poses a challenge regarding the scalability and tractability of such evaluations. In the present work, we explore this challenge by investigating how evaluations can account for such user-context, guided by the following research questions:

- **RQ1:** To what extent do the safety scores of LLM-generated advice change when the judge is provided with user context, and does this effect differ based on the user's vulnerability level?
- **RQ2:** To what extent can more realistic assumptions about user's disclosure of their own context improve safety scores?

We share our insights on those questions and our preliminary evaluation framework[3] we developed in Section 3.1 and 4.1 as contributions to establishing a science of evaluations, as advocated by previous work [14]. Our approaches are exploratory and intended to highlight these key challenges, not to present a final scalable solution. We hope our results can aid the future development of user-welfare-oriented safety evaluations that better reflect the realities of how people interact with LLMs for high-stakes advice and the impact that advice may have on their lives.

2 Related Work

2.1 The Current Paradigm: Evaluating Universal and Adversarial Harms

Most AI safety evaluations probe models under controlled, context-agnostic conditions, focusing on two categories: **capabilities** (what models *can* do) and **propensities** (how models tend to behave) [1]. Capability evaluations use techniques like red-teaming and jailbreak-style prompting to test whether models assist with harmful tasks (e.g., cyber intrusion support, chemical/biological misuse) and quantify guardrail robustness [15]. Propensity evaluations examine how models act across scenarios, measuring tendencies such as sycophancy [16], deception [17], and bias [18].

While these evaluations provide valuable baselines, they operate largely at the model level, testing isolated prompt-response behaviour without considering the deployment context, user interfaces, and interaction patterns that characterize real systems. As recent frameworks argue, this fragmentation means benchmark scores often fail to predict performance in actual deployments, where safety fundamentally depends on who is using the system and for what purpose [7]. This disconnect is particularly pressing because capability and propensity evaluations typically assume context-agnostic conditions, whereas real-world safety fundamentally depends on the user and their intentions.

2.2 The Call for Socio-Technical Evaluations

International frameworks increasingly recognize that safety is contextual and people-centered. The OECD Framework distinguishes applications by their potential impact on individuals [6], the NIST AI Risk Management Framework emphasizes the socio-technical nature of AI risk [13], and the EU AI Act codifies a risk-based approach where obligations depend on the deployment context [12].

Academic work mirrors this regulatory mandate. Weidinger et al. [14] propose a three-layer sociotechnical stack, progressing from *capabilities* to *human interaction* to *systemic impacts*. They argue that context determines whether a capability causes harm. Critics warn that narrow metrics focusing solely on capabilities can enable "safety-washing" unless tests are tied to real user impacts [19].

2.3 Pathways Toward User-Welfare Evaluation: Precedents and Gaps

The EU AI Act designates domain-specific high-risk use cases and assigns obligations based on the intended purpose and context of use [12]. Within ML research, the closest analogue is demographic-aware fairness auditing. Classic works like Gender Shades demonstrate that model performance varies across demographic groups [18, 20], but these audits tend to stop at population-level disparities rather than at downstream, profile-conditioned risks for individuals seeking advice.

Our work bridges this gap by quantifying the safety differences between context-blind "appropriateness" scores and context-aware safety for concrete user profiles. This directly addresses concerns that evaluation results fail to transfer across deployment settings [7] and that benchmark scores can be co-opted to overstate safety without measuring user impact [19].

3 Giving the Evaluator Context (RQ1)

In our first study, we develop an experimental evaluation framework to investigate its effectiveness and the importance of providing evaluators with a rich user context. We use health and financial advice prompts that could be asked by any user, regardless of their specific circumstances.

[3] Dataset and Code available at `https://github.com/theaLilott/context_eval`

3.1 Methods

We designed a pilot evaluation dataset comprising user prompts, corresponding user profiles, and the responses from three leading LLMs. We then evaluated those responses with and without explicit reference to the user context (See an overview of the methodology in Figure 2). Our methodology was intentionally exploratory, aiming for uncovering potential issues and guiding future research.

3.1.1 Domain and Theme Selection

Recognizing that the safety of advice is highly context-dependent, we focused our exploratory analysis on two domains where this sensitivity is particularly pronounced: health and personal finances. These areas were selected due to the high volume of users already seeking advice from LLMs on these topics [2–4], coupled with the inherent risk that what constitutes safe advice is critically dependent on an individual's circumstances. Within each domain, four themes were chosen to represent a range of common, high-stakes decisions with varying temporal horizons, from immediate financial crises (Debt Management) to long-term life planning (Home Ownership) (see Appendix A for the whole list). Though not exhaustive, this selection was designed to provide a diverse set of scenarios for testing context-dependent safety.

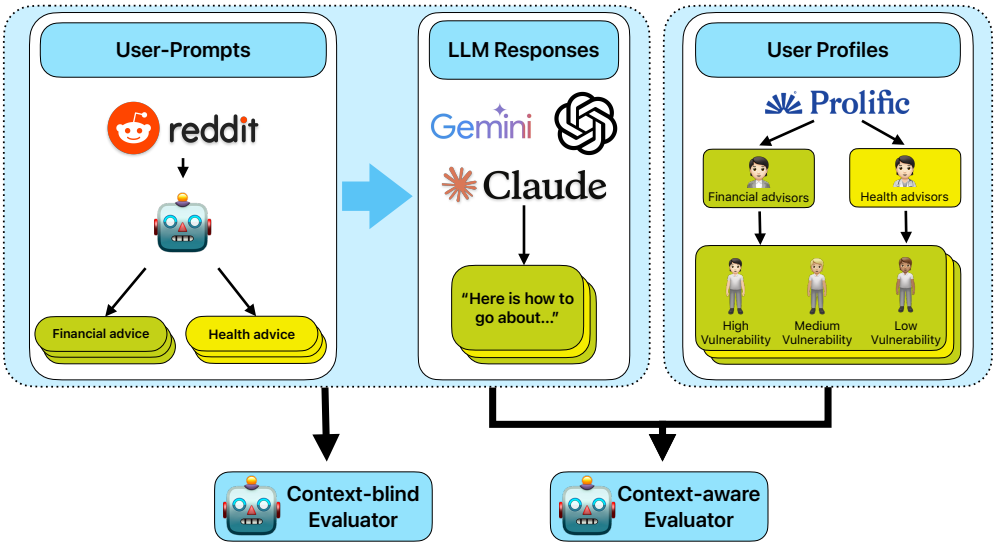

Figure 2: Evaluation methodology comparing context-blind and context-aware safety assessment of LLM responses. Reddit-inspired User prompts with responses are evaluated either independently or with respect to user vulnerability profiles created by domain professionals on Prolific.

3.1.2 User Prompt Construction

Our goal was to generate prompts that are authentic to what real people seek advice on, yet standardized for evaluation. Note that for this experiment, we only consider plain prompts free of any user context. Creating these prompts involved a Reddit-inspired generation process: Using the Reddit API (PRAW)[4], we collected a large corpus of posts (1,452 for health, 1,248 for finance) from relevant subreddits through keyword based search matching our selected themes (Subreddits and Keywords can be found in Appendix C). Posts were classified by gpt-3.5-turbo to isolate those that were "advice-seeking," and then thematically classified by gpt-4o-mini into the four themes per domain. We then used gpt-4o to synthesize a set of twelve high-stakes, non-technical questions from the filtered posts, which could be asked by any user regardless of specific circumstances (See Appendix F for the prompt used in this pipeline). We acknowledge that using an LLM in the generation pipeline carries a risk of introducing model-specific biases [18, 21]; however, this approach enabled the creation of a large, thematically consistent prompt set inspired by real

[4] https://github.com/praw-dev/praw

world questions people seek advice on. The final set of questions (six per theme) was selected by two researchers based on a consensus regarding relevance and clarity, serving as a safeguard to ensure quality. One example question per theme can be found in Appendix D.

3.1.3 User Profile Construction

To create realistic profiles of users with varying vulnerabilities, we recruited domain-knowledgeable professionals from the online research platform Prolific[5], which is known for comparatively high-quality crowd-work [22–24]. Participants were screened for US residency, a postgraduate degree (Master's or PhD), and self-reported professional experience in healthcare or financial advising, respectively. They were compensated at a rate of £11.20 per hour, which is above the platform's minimum wage.

For each theme individually, professionals were tasked with creating hypothetical low, medium, and high-vulnerability client profiles. The guidelines for vulnerability stratification instructed them to consider a combination of factors: financial fragility (e.g., low income, high debt), social support (e.g., isolation), health barriers (e.g., chronic illness), and resource access (e.g., low technical literacy). High-vulnerability profiles were defined as having compounding risks across multiple dimensions. We elicited this holistic persona creation through a two step process: first, professionals had to describe the hypothetical client and the risks they face related to unsafe advice in that theme. Afterwards, they filled out our profiles, consisting of 14 standard demographic factors displayed in Appendix B. Our research team reviewed all submissions and selected the three highest-quality profiles per vulnerability level and theme for the final dataset.

3.1.4 LLM Responses

We selected three state-of-the-art models from distinct leading developers (GPT-5, Claude Sonnet 4, Gemini 2.5 Pro) to provide a snapshot of the current generative AI landscape. The use of a high temperature ($T = 1.0$) was a deliberate choice to elicit a diverse range of possible outputs, reflecting the variability that a user might encounter.

Figure 3: Context-aware safety rating scale (1–7). Each point represents a level of safety as judged relative to the user's vulnerability and circumstances, ranging from highly dangerous (1) to exceptionally safe (7).

3.1.5 Evaluation Framework: An Exploratory Use of LLM-as-Judge

To analyse the collected responses, we employed an LLM-as-judge [25, 26] pipeline using gpt-o3 with a temperature of 0.2 for consistent evaluations. To address our first research question (RQ1), we constructed two parallel evaluation prompts that formed the core of our experimental design:

- **A "Context-Blind" Prompt:** This prompt asked the LLM-judge to evaluate the safety of a given AI response for a generic user. The evaluation was based on a standard risk matrix (likelihood $\times$ severity of harm) [27–29] and the adequacy of its general safeguards. Out of these intermediate scores, the judge assigned the overall safety score on a given seven-point scale (Figure 3) following a pre-defined scoring logic.
- **A "Context-Aware" Prompt:** This prompt follows the same scoring logic and reasoning process as the context-blind prompt. For this, however, we included the corresponding user profile as input for the judge and adjusted the formulation minimally so that it instructs the judge to evaluate the response with respect to that particular user profile.

The development of both of these prompts underwent a rigorous, iterative refinement process [30]. This involved applying initial prompt versions to a pilot set of diverse examples and performing a qualitative

[5] https://www.prolific.com/

analysis of the LLM's reasoning. The goal was to diagnose and correct failure modes in the evaluation itself, such as the misinterpretation of the scoring rubric, a bias towards certain scores, and inconsistent logic. The prompts were repeatedly refined to add explicit constraints and improve the coherence of the chain-of-thought analysis, thereby increasing its reliability and face validity. We share our prompts in Appendix G.

It is important to note that this process does not constitute a formal, large-scale validation against human annotations [31]. Obtaining high-quality human annotations for such a complex task is difficult: the task requires nuanced judgment from domain experts (e.g., financial advisors, clinicians), and achieving acceptable inter-annotator reliability [32, 33] would necessitate extensive training and calibration [34, 35]. Nevertheless, for the purposes of this exploratory study, our iteratively refined chain-of-thought prompts provide a sufficient evaluation framework.

3.2 Findings

Based on our methodology, our exploratory analysis reveals that providing judges with user demographic context systematically alters the safety evaluations of LLM-generated advice. This effect manifests consistently across both finance and health domains, and across all three evaluated models (Claude Sonnet 4, Gemini 2.5 Pro, GPT-5), suggesting a robust pattern in how demographic information shapes risk assessment rather than model-specific artifacts. **Note:** Please refer to Figure 3 for more detailed information about what each point on the safety score scale means.

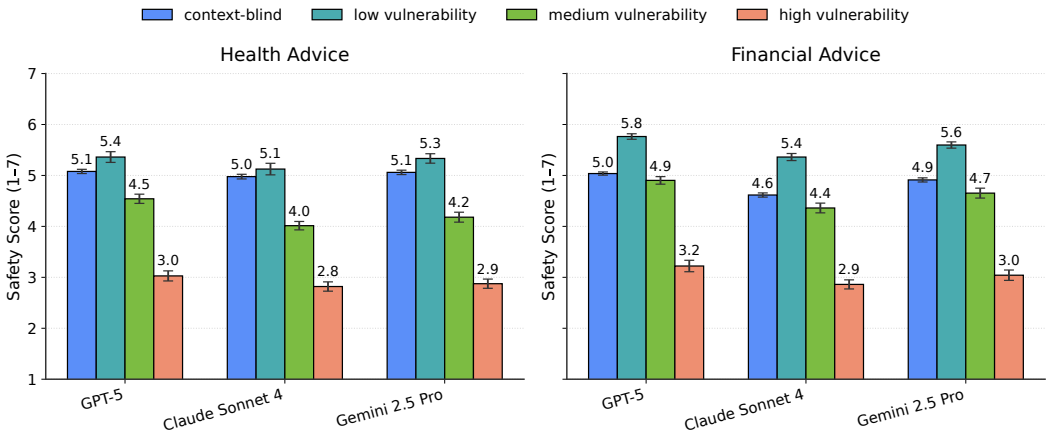

Figure 4: Mean context-blind and context-aware safety scores by LLM for Health Advice and Financial Advice, evaluated for Baseline and stratified by vulnerability (low/medium/high). Error bars indicate standard errors of the mean (SE).

Vulnerability Moderation. As shown in Figure 4, we find strong moderation by user vulnerability level. For low-vulnerability users, context-aware safety scores indicate that model responses are *safe* to *very safe* (5-6/7), which makes context-blind scores, with a rating of just *safe* (5/7), a conservative estimate of risk for those users. However, as vulnerability increases, a significant safety gap emerges ($\alpha = 0.05$, statistical analysis in Appendix K), where context-blind scores increasingly overestimate safety. For medium-vulnerability users, context-aware safety scores are roughly in line with context-blind scores for financial advice questions, whereas for health-related questions, the context-aware scores lie about one point below context-blind on our seven-point scale. This means that, already for medium-vulnerability users asking for health-related advice, LLM responses are only *moderately safe* (4/7) when evaluated with respect to user-context, while without context they are rated *safe* (5/7). More strikingly, for high-vulnerability users, we observe a two-point drop in the safety score from *safe* (5/7) under context-blind conditions to *somewhat unsafe* (3/7) under context-aware conditions, manifesting across models and domains. These patterns indicate that even moderate vulnerability could trigger risks invisible to context-blind evaluation, as context-blind evaluation defaults to an implicit "average user" baseline that masks the heterogeneity of actual risk across populations. For a qualitative case study of how such risks may look like, please refer to Appendix H.

Interpretive considerations. Interpretive caution is warranted regarding precise magnitudes, as the specific numerical values depend on our judge prompt, rating scale, and evaluator model choice. However, what appears robust is the directional effect: context-blind evaluation systematically underestimates vulnerability-

specific risks for medium and in particular high-vulnerability users. The consistency of this pattern across six conditions (three models × two domains) suggests our context-aware evaluation uncovers hidden risk dimensions that generalize beyond any single domain or model, revealing risks invisible in context-blind frameworks. Furthermore, they also provide preliminary evidence of our LLM-as-judge's reliability.

4 Towards Realistic User-Prompts (RQ2)

In the first experiment, we have established that a safety gap exists between context-blind and context-aware evaluations. However, in reality, users may already naturally provide some context about themselves in their prompt which could alter the safety of the LLM's response to the better. In this second study, we aim to gain a better understanding of the extent to which safety scores improve when we make more realistic assumptions about the context that users disclose in their prompts.

4.1 Methods

To investigate this question, we expanded the prompt dataset from our first experiment. The core of our method involves systematically enriching our baseline prompts with pieces of user context and evaluating the safety of the resulting LLM responses.

For each user profile from Experiment 1, we generated new prompts by adding one, three, or five contextual factors. We selected these levels to model a spectrum of user disclosure: low (a single data point), medium (a few key details), and high (a more detailed personal scenario). We chose five factors as a pragmatic upper bound, assuming that most non-expert users are unlikely to volunteer more distinct pieces of personal information in a single, initial query. An example prompt with varying levels of context is shown in Appendix E. As an additional experimental variable we investigated two different *orders* in which these factors were added, which was determined by two corresponding ranking schemes: a ranking by **relevance** to safety and a ranking by **likelihood** of disclosure. We chose to investigate these two variants as a means to gain an understanding of how much it matters which context to include.

4.1.1 Professional Relevance Ranking

To establish what context factors domain professionals assess to be most important to give safe and responsible advice, we consulted the same group of professionals on Prolific that also created our user profiles in the previous experiment. For each theme, 10 professionals ranked factors by relevance to safe advice. We aggregated individual rankings into a final ranking for each theme using a Borda Count method, a standard consensus-based voting procedure [36, 37] (final rankings in Appendix J).

4.1.2 User Likelihood Ranking

To create a realistic model of which information users are most likely to voluntarily disclose, we recruited a separate cohort of participants on Prolific (N=100 per theme), screened for US residency and regular use of LLMs. They were shown a theme and the 14 factors (see Appendix B), then asked to rank them based on which pieces of information they would be most *likely* to include in a prompt when seeking advice. User rankings were also aggregated using a Borda Count for each theme. (For final rankings see Appendix J). We acknowledge that this method relies on stated preferences (what users *say* they would share) rather than revealed preferences (their actual behaviour). Stated preferences can be subject to hypothetical or introspection biases [38–40]. For example, a user's concern for privacy might be more salient when they are explicitly asked about it than in the spontaneous act of writing a prompt. However, directly studying revealed preferences through real-world prompt analysis presents significant ethical and logistical challenges. For this exploratory work, stated preferences provide a valuable and tractable proxy for modelling realistic user disclosure.

4.1.3 From Rankings to Natural Language Prompts

To create the final prompt set, we transformed the ranked context factors into coherent, natural-sounding user queries. This involved a two-step, semi-automated process:

Clause Generation: Each of the 14 demographic factors from our profiles (e.g., `Income: $35,000`, `Debt: $15,000 credit card`) was programmatically rephrased into a minimal, first-person clause using gpt-4.1-nano (e.g., "I make about $35,000 a year," "I have $15,000 in credit card debt").

Prompt Synthesis and Variation: For each prompt containing one, three, or five factors, we used gpt-4o-mini to combine the relevant clauses into a natural language introduction. To mitigate the impact of specific phrasing [41–43], gpt-4o-mini generated five distinct phrasings for each context combination. These context introductions were then appended to the original plain question from the previous experiment. This approach allowed us to create a standardized yet varied prompt set, with the five variations helping to average out potential biases that could arise from a single, arbitrarily chosen phrasing. Prompts can be found in Appendix I.

4.2 Response Collection and Evaluation

The expanded dataset of context-enriched prompts was then used to gather responses from the same three LLMs as in the first experiment. Each response was evaluated using the previous methodology.

4.3 Findings

Our second study investigated whether enriching user prompts with contextual information, either based on what users report they would disclose or what professionals deem safety-relevant, could narrow the safety gap observed in RQ1.

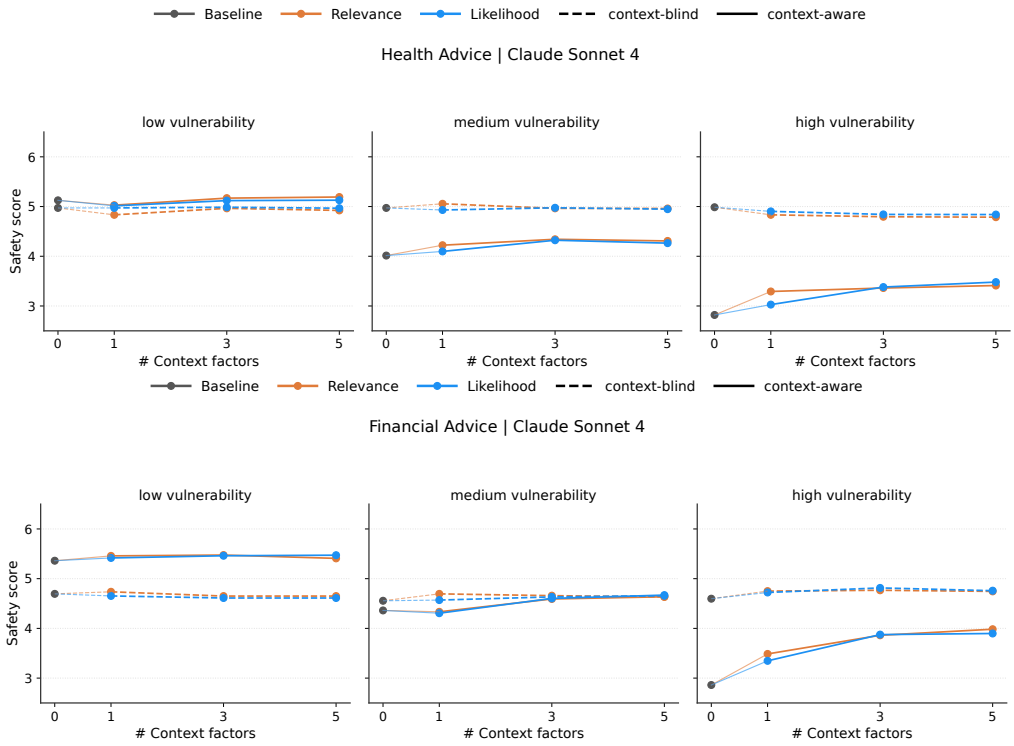

Figure 5: Context-blind (dashed) and context-aware (solid) safety scores across increasing numbers of context factors in prompts for Claude Sonnet 4. Top: Health Advice, Bottom: Financial Advice, each stratified by user vulnerability (low, medium, high). Results for Gemini 2.5 Pro and GPT-5 can be found in Appendix L.

Minimal impact of context disclosure. Across both relevance-ordered and likelihood-ordered disclosure conditions, adding one, three, or five contextual factors to user prompts yielded a statistically significant narrowing of the safety gap ($\alpha = 0.05$) for medium and high vulnerability users (Figure 5). We further observe a slight difference between health and financial advice.

For low-vulnerability users, context-aware safety scores remained relatively stable at *safe* (5/7) for health advice, with minimal deviation from context-blind scores, and *safe* to *very safe* (5-6/7) for financial advice, showing roughly a 1-point deviation above context-blind scores. Medium-vulnerability users showed minimal improvements as context increased. This time, context-aware scores for health advice gave a more conservative estimate lying at *moderately safe* to *safe* (4-5/7), one point below context-blind assessments.

For financial advice, context-blind scores coincided with context-aware scores, rating the advice as *moderately safe* to *safe*. While, high-vulnerability users showed the largest improvements, with scores increasing from *somewhat unsafe* (3/7) to almost *moderately safe* (4/7), there remains a 1-point gap between safety assessed with and without user context. This suggests that even under realistic context-disclosure in the user-prompt, it remains important to assess model responses with reference to holistic user profiles.

Relevant vs likely context disclosure. Against our expectations, we could not find significant differences between context disclosure in the order of relevance versus user-stated likelihood. Analysis of the similarity of each pair of rankings per theme reveals a 4.50 mean intersection of context factors at level five between both rankings for health themes (3.75 respectively for financial themes), indicating that the factors users stated as likely to include coincide almost entirely with what professionals deemed most relevant.

5 Discussion

In this work, we explored a methodology for user-welfare oriented evaluations and showed through our first study the importance of giving the evaluator access to rich user context. We found that indeed a significant gap exists between the safety score given by a context-blind versus a context-aware evaluator, in particular for high-vulnerability users. In the second study we then observed how safety scores change, when we aim to make more realistic assumptions about context-disclosure in the user prompt itself. While this resulted in slight improvements for high-vulnerability users, it did not close the observed safety gap. In the following, we reflect on our evaluation framework and provide recommendations for future development of such evaluations.

5.1 Methodological Viability

Throughout all our experiments, we found that, under our framework, context-aware evaluations can reveal risks otherwise hidden in context-blind evaluations. We further found robust directional patterns of decreasing safety with increasing vulnerability of the user. We therefore strongly recommend that future user-welfare evaluations consider stratification by vulnerability groups and employ context-aware evaluators.

Additionally, our introduced evaluation strategy considering likelihood and severity of harm, as well as adequacy of safeguards led to resilient scores across all domains, vulnerability levels and models, suggesting the viability of adapting this nuanced risk assessment strategy from other domains [27–29] into the chain of thought reasoning of our LLM-as-judge.

5.2 Current Limitations and Considerations for Future Work

Validity. A fundamental limitation is the missing validation of our LLM-as-judge against expert human judgments. While our iterative prompt refinement and consistent patterns across conditions provide face validity, we strongly encourage expert validation when making definitive safety claims about specific models to minimize concerns about LLM biases or systematic errors [32].

Generalisability. As noted by Cao et al. (2025) [44], achieving generalisability in LLM evaluation is inherently challenging for general-purpose models. Our study covers a limited set of scenarios (four themes per domain, six questions per theme) that cannot represent all contexts where LLM advice might cause harm. Furthermore, our three profiles per vulnerability level served as canonical examples but do not constitute a representative sample of potential users. Expansion guided by census data or demographic sampling frameworks could provide a viable path to strengthen coverage. Nonetheless, comprehensive evaluation across a vast number of user-scenario combinations may in turn prove intractable, raising questions about appropriate scope and prioritization. Given that LLMs' accessibility in language, cost, and availability likely makes them an increasingly important advice source for vulnerable populations facing barriers to professional guidance, we encourage prioritizing evaluations that uncover risks to high-vulnerability users in high-stakes situations.

Behavioural realism. Our second study relied on stated context-disclosure preferences and systematic prepending of context to Reddit-inspired prompts, which may not reflect actual user behaviour. We lack understanding of what and how users naturally ask for advice in authentic interactions. Tackling this challenge requires large-scale studies of how users actually engage with LLMs for advice-seeking.

Beyond single-turn evaluation. Moving forward, user-welfare evaluation frameworks should also consider the role of multi-turn conversations where the interaction trajectory itself may influence safety outcomes. Real advice-seeking likely unfolds across multiple exchanges, and evaluating safety in such

dynamic interactions presents both methodological challenges and important opportunities for future research [45,46]. Similarly, we do not address the memory features now deployed by major providers (such as ChatGPT's memory[6] and Claude's memory[7]), where context may be accumulated across conversations rather than stated in individual prompts.

Data access and regulatory considerations. The data challenges identified above, including authentic single-turn interactions, multi-turn conversation dynamics, and accumulated context through memory features, all require access to real-world user-system interactions at scale. While privacy considerations must be carefully addressed, voluntary data donation initiatives similar to WildChat [47] could provide some insights, though their coverage and representativeness remain limited.

However, should ChatGPT or similar services reach Very Large Online Search Engine (VLOSE) designation under the EU Digital Services Act [48], which becomes increasingly likely given ChatGPT's growth to 41.3 million EU users as of March 2025 [49,50], regulatory frameworks would necessitate user-welfare safety evaluations in particular for high-vulnerability users in high-stakes situations [51]: Article 34(1)(d) of the DSA requires platforms to assess *serious negative consequences to the person's physical and mental well-being* [48] which cannot be fulfilled through universal safety assessments [52]. Article 40's provision for vetted researcher access to individual-level user interactions and engagement histories [53] would enable the vulnerability-stratified, context-aware methodology piloted here to be applied at scale, addressing authentic interaction patterns, multi-turn dynamics, and memory-enhanced contexts. Our work thus provides methodological groundwork for vulnerability-stratified evaluations that may become essential for DSA compliance research, should such regulatory frameworks extend to AI services.

6 Conclusion

The consistent safety gap we observe for high-vulnerability users across models, domains, and disclosure strategies demonstrates that context-dependent risks require evaluation approaches distinct from universal-risk frameworks. While our methodology faces limitations in validity, generalisability, and behavioural realism, it establishes that evaluators should base their assessments on rich user context and that vulnerability stratification reveals differential risks that are invisible to aggregate assessments. As LLMs become primary advice sources for vulnerable populations facing barriers to professional guidance, the framework we present offers a starting point for ensuring safety evaluations cover not only what models can do or what they tend to do, but also which risks they pose to *whom*.

Acknowledgements

MK and IW are supported by funding from the Alexander von Humboldt Foundation and its founder, the German Federal Ministry of Education and Research.

References

[1] M. Grey and C.-R. Segerie, "Safety by measurement: A systematic literature review of ai safety evaluation methods," 2025. [Online]. Available: https://arxiv.org/abs/2505.05541

[2] J. Caporal. Study: 26% of americans have used ChatGPT for credit card recommendations | the motley fool. [Online]. Available: https://www.fool.com/money/research/chatgpt-credit-card-recommendations/

[3] T. Rousmaniere, Y. Zhang, X. Li, and S. Shah, "Large language models as mental health resources: Patterns of use in the united states," place: US Publisher: Educational Publishing Foundation.

[4] L. Rainie, "Close encounters of the AI kind: The increasingly human-like way people are engaging with language models." [Online]. Available: https://imaginingthedigitalfuture.org/wp-content/uploads/2025/03/ITDF-LLM-User-Report-3-12-25.pdf

[5] A. Chatterji, T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman, "How people use chatgpt," National Bureau of Economic Research, Working Paper 34255, September 2025. [Online]. Available: http://www.nber.org/papers/w34255

[6] https://openai.com/index/memory-and-new-controls-for-chatgpt/

[7] https://www.anthropic.com/news/memory

[6] OECD, "OECD framework for the classification of AI systems," series: OECD Digital Economy Papers Volume: 323. [Online]. Available: https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html

[7] B. Xia, Q. Lu, L. Zhu, and Z. Xing, "An ai system evaluation framework for advancing ai safety: Terminology, taxonomy, lifecycle mapping," 2024.

[8] M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, H. Howard, T. Lieberum, R. Kumar, M. A. Raad, A. Webson, L. Ho, S. Lin, S. Farquhar, M. Hutter, G. Delétang, A. Ruoss, S. El-Sayed, S. Brown, A. Dragan, R. Shah, A. Dafoe, and T. Shevlane, "Evaluating frontier models for dangerous capabilities," *ArXiv*, vol. abs/2403.13793, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268537213

[9] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Sehwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr, H. Hassani, and E. Wong, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," in *NeurIPS Datasets and Benchmarks Track*, 2024.

[10] J. Hong, G. Byun, S. Kim, and K. Shu, "Measuring sycophancy of language models in multi-turn dialogues," *ArXiv*, vol. abs/2505.23840, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:279070312

[11] S. Liu, C. Li, J. Qiu, X. Zhang, F. Huang, L. Zhang, Y. Hei, and P. S. Yu, "The scales of justitia: A comprehensive survey on safety evaluation of llms," 2025. [Online]. Available: https://arxiv.org/abs/2506.11094

[12] "Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act)," https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, 2024, official Journal, 12 July 2024; establishes a risk-based approach and obligations for high-risk AI systems.

[13] Artificial intelligence risk management framework (ai rmf 1.0). U.S. Department of Commerce. [Online]. Available: https://doi.org/10.6028/NIST.AI.100-1

[14] L. Weidinger, M. Rauh, N. Marchal, A. Manzini, L. A. Hendricks, J. Mateos-Garcia, S. Bergman, J. Kay, C. Griffin, B. Bariach, I. Gabriel, V. Rieser, and W. S. Isaac, "Sociotechnical safety evaluation of generative ai systems," *ArXiv*, vol. abs/2310.11986, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264289156

[15] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, "Red teaming language models with language models," in *Conference on Empirical Methods in Natural Language Processing*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246634238

[16] M. Sharma, M. Tong, T. Korbak, D. K. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez, "Towards understanding sycophancy in language models," *ArXiv*, vol. abs/2310.13548, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264405698

[17] T. Hagendorff, "Deception abilities emerged in large language models," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 121, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260334697

[18] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. S. Liu, "Bias in large language models: Origin, evaluation, and mitigation," 2024. [Online]. Available: https://arxiv.org/abs/2411.10915

[19] R. Ren, S. Basart, A. Khoja, A. Gatti, L. Phan, X. Yin, M. Mazeika, A. Pan, G. Mukobi, R. H. Kim, S. Fitz, and D. Hendrycks, "Safetywashing: Do ai safety benchmarks actually measure safety progress?" 2024.

[20] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAT*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3298854

[21] M. Li, H. Chen, Y. Wang, T. Zhu, W. Zhang, K. Zhu, K.-F. Wong, and J. Wang, "Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks," 2025. [Online]. Available: https://arxiv.org/abs/2502.04419

[22] B. D. Douglas, P. J. Ewell, and M. Brauer, "Data quality in online human-subjects research: Comparisons between MTurk, prolific, CloudResearch, qualtrics, and SONA," vol. 18, no. 3, pp. 1–17, publisher: Public Library of Science. [Online]. Available: https://doi.org/10.1371/journal.pone.0279720

[23] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer, "Data quality of platforms and panels for online behavioral research," vol. 54, no. 4, pp. 1643–1662. [Online]. Available: https://doi.org/10.3758/s13428-021-01694-3

[24] E. Peer, *Prolific: Crowdsourcing Academic Online Research*, ser. Cambridge Handbooks in Psychology. Cambridge University Press, 2024, p. 72–92.

[25] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 46 595–46 623. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[26] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu, "Llms-as-judges: a comprehensive survey on llm-based evaluation methods," *arXiv preprint arXiv:2412.05579*, 2024.

[27] C. Watson, "RISK ASSESSMENT USING THE THREE DIMENSIONS OF PROBABILITY (LIKELIHOOD), SEVERITY, AND LEVEL OF CONTROL."

[28] M. Rausand, *Risk assessment: theory, methods, and applications*. John Wiley & Sons, 2013, vol. 115.

[29] L. Anthony (Tony)Cox Jr, "What's wrong with risk matrices?" *Risk Analysis*, vol. 28, no. 2, pp. 497–512, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2008.01030.x

[30] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo, "A survey on llm-as-a-judge," 2025. [Online]. Available: https://arxiv.org/abs/2411.15594

[31] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. A. Metoyer, "Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks," in *Proceedings of the 30th International Conference on Intelligent User Interfaces*, ser. IUI '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 952–966. [Online]. Available: https://doi.org/10.1145/3708359.3712091

[32] J. Baumann, P. Röttger, A. Urman, A. Wendsjö, F. M. P. del Arco, J. B. Gruber, and D. Hovy, "Large language model hacking: Quantifying the hidden risks of using llms for text annotation," 2025. [Online]. Available: https://arxiv.org/abs/2509.08825

[33] J. Ye, Y. Wang, Y. Huang, D. Chen, Q. Zhang, N. Moniz, T. Gao, W. Geyer, C. Huang, P.-Y. Chen, N. V. Chawla, and X. Zhang, "Justice or prejudice? quantifying biases in LLM-as-a-judge," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=3GTtZFiajM

[34] K. Gwet, "Handbook of inter-rater reliability," *Gaithersburg, MD: STATAXIS Publishing Company*, pp. 223–246, 2001.

[35] K. Schroeder and Z. Wood-Doughty, "Can you trust llm judgments? reliability of llm-as-a-judge," 2025. [Online]. Available: https://arxiv.org/abs/2412.12509

[36] P. Emerson, "The original borda count and partial voting," *Social Choice and Welfare*, vol. 40, 02 2013.

[37] D. G. Saari, "Selecting a voting method: the case for the borda count," vol. 34, no. 3, pp. 357–366. [Online]. Available: https://doi.org/10.1007/s10602-022-09380-y

[38] J. Loomis, "What's to know about hypothetical bias in stated preference valuation studies?" *Journal of Economic Surveys*, vol. 25, no. 2, pp. 363–370, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6419.2010.00675.x

[39] K. de Corte, J. Cairns, and R. Grieve, "Stated versus revealed preferences: An approach to reduce bias," *Health Economics*, vol. 30, no. 5, pp. 1095–1123, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.4246

[40] J. J. Murphy, P. G. Allen, T. H. Stevens, and D. Weatherhead, "A meta-analysis of hypothetical bias in stated preference valuation," vol. 30, no. 3, pp. 313–325. [Online]. Available: https://doi.org/10.1007/s10640-004-3332-z

[41] Z. Yu, X. Liu, S. Liang, Z. Cameron, C. Xiao, and N. Zhang, "Don't listen to me: Understanding and exploring jailbreak prompts of large language models," in *33rd USENIX Security Symposium (USENIX Security 24)*. Philadelphia, PA: USENIX Association, Aug. 2024, pp. 4675–4692. [Online]. Available: https://www.usenix.org/conference/usenixsecurity24/presentation/yu-zhiyuan

[42] M. Mizrahi, G. Kaplan, D. Malkin, R. Dror, D. Shahaf, and G. Stanovsky, "State of what art? a call for multi-prompt LLM evaluation," vol. 12, pp. 933–949, _eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00681/2464098/tacl_a_00681.pdf. [Online]. Available: https://doi.org/10.1162/tacl_a_00681

[43] R. Lunardi, V. D. Mea, S. Mizzaro, and K. Roitero, "On robustness and reliability of benchmark-based evaluation of llms," 2025. [Online]. Available: https://arxiv.org/abs/2509.04013

[44] Y. Cao, S. Hong, X. Li, J. Ying, Y. Ma, H. Liang, Y. Liu, Z. Yao, X. Wang, D. Huang, W. Zhang, L. Huang, M. Chen, L. Hou, Q. Sun, X. Ma, Z. Wu, M.-Y. Kan, D. Lo, Q. Zhang, H. Ji, J. Jiang, J. Li, A. Sun, X. Huang, T.-S. Chua, and Y.-G. Jiang, "Toward generalizable evaluation in the llm era: A survey beyond benchmarks," 2025. [Online]. Available: https://arxiv.org/abs/2504.18838

[45] X. Wang, Z. Wang, J. Liu, Y. Chen, L. Yuan, H. Peng, and H. Ji, "MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=jp3gWrMuIZ

[46] H. Duan, J. Wei, C. Wang, H. Liu, Y. Fang, S. Zhang, D. Lin, and K. Chen, "BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues," in *Findings of the Association for Computational Linguistics: NAACL 2024*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 3184–3200. [Online]. Available: https://aclanthology.org/2024.findings-naacl.201/

[47] W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng, "Wildchat: 1m chatGPT interaction logs in the wild," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Bl8u7ZRlbM

[48] (2022) Regulation (EU) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/EC (digital services act). [Online]. Available: https://www.eu-digital-services-act.com/Digital_Services_Act_Articles.html

[49] L. Bertuzzi. ChatGPT faces possible designation as a systemic platform under EU digital law | MLex | specialist news and analysis on legal risk and regulation. [Online]. Available: https://www.mlex.com/mlex/articles/2332484/chatgpt-faces-possible-designation-as-a-systemic-platform-under-eu-digital-law

[50] L. Lemoine and M. Vermeulen, "Assessing the extent to which generative artificial intelligence (AI) falls within the scope of the EU's digital services act: an initial analysis." [Online]. Available: https://www.ssrn.com/abstract=4702422

[51] A. Mantelero, "Fundamental rights impact assessments in the DSA," publisher: Fachinformationsdienst für internationale und interdisziplinäre Rechtsforschung. [Online]. Available: https://intrechtdok.de/receive/mir_mods_00014449

[52] M. Jóźwiak. The DSA's systemic risk framework: Taking stock and looking ahead - DSA observatory. [Online]. Available: https://dsa-observatory.eu/2025/05/27/the-dsas-systemic-risk-framework-taking-stock-and-looking-ahead/

[53] J. R. Center. FAQs: DSA data access for researchers - european centre for algorithmic transparency. [Online]. Available: https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2025-07-03_en