Can't LLMs do that? Supporting Third-Party Audits under the DSA: Exploring Large Language Models for Systemic Risk Evaluation of the Digital Services Act in an Interdisciplinary Setting

Sekwenz, Marie-Therese; Gsenger, Rita; Stocker, Volker; Görnemann, Esther; Talypova, Dinara; Parkin, Simon; Greminger, Lea; Smaragdakis, Georgios

DOI
10.1145/3707640.3731929

Publication date
2025

Document Version
Final published version

Published in
CHIWORK '25 Adjunct: Adjunct Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work

Can't LLMs do that? Supporting Third-Party Audits under the DSA: Exploring Large Language Models for Systemic Risk Evaluation of the Digital Services Act in an Interdisciplinary Setting

Marie-Therese Sekwenz
Multi-Actor Systems/ Organisation and Governance / AI Futures Lab
TU Delft
Delft, Netherlands
M.T.Sekwenz@tudelft.nl

Rita Gsenger
Weizenbaum Institut
Berlin, Germany
rita.gsenger@hu-berlin.de

Volker Stocker
Weizenbaum Institut
Berlin, Germany
volker.stocker@weizenbaum-institut.de

Esther Görnemann
Weizenbaum Institut
Berlin, Germany
esther.goernemann@weizenbaum-institut.de

Dinara Talypova
IT:U
Linz, Austria
dinara.talypova@it-u.at

Simon Parkin
Multi-Actor Systems/ Organisation and Governance / Cyber Security
TU Delft
Delft, Netherlands
s.e.parkin@tudelft.nl

Lea Greminger
Weizenbaum Institut
Berlin, Germany
lea.greminger@weizenbaum-institut.de

Georgios Smaragdakis
Electrical Engineering, Mathematics and Computer Science /Intelligent Systems / Cyber Security
TU Delft
Delft, Netherlands
g.smaragdakis@tudelft.nl

Abstract

This paper investigates the feasibility and potential role of using Large Language Models (LLMs) to support systemic risk audits under the European Union's Digital Services Act (DSA). It examines how automated tools can enhance the work of DSA auditors and other ecosystem actors by enabling scalable, explainable, and legally grounded content analysis. An interdisciplinary expert workshop with twelve participants from legal, technical, and social science backgrounds explored prompting strategies for LLM-assisted auditing. Thematic analysis of the sessions identified key challenges and design considerations, including prompt engineering, model interpretability, legal alignment, and user empowerment. Findings highlight the potential of LLMs to improve annotation workflows and expand audit scale, while underscoring the continued importance of human oversight, iterative testing, and cross-disciplinary collaboration. This study offers practical insights for integrating AI tools into auditing processes and contributes to emerging methodologies for operationalizing systemic risk evaluations under the DSA.

CCS Concepts

• **Social and professional topics** → **Computing / technology policy**; • **Computing methodologies** → **Natural language processing**; • **Information systems** → **Language models**; • **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

Large Language Models, Digital Services Act, Online Platform Auditing, Systemic Risk, Content Moderation, Human-AI Collaboration.

ACM Reference Format:
Marie-Therese Sekwenz, Rita Gsenger, Volker Stocker, Esther Görnemann, Dinara Talypova, Simon Parkin, Lea Greminger, and Georgios Smaragdakis. 2025. Can't LLMs do that? Supporting Third-Party Audits under the DSA: Exploring Large Language Models for Systemic Risk Evaluation of the Digital Services Act in an Interdisciplinary Setting. In *CHIWORK '25 Adjunct: Adjunct Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '25 Adjunct), June 23–25, 2025, Amsterdam, Netherlands.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3707640.3731929

CHIWORK '25 Adjunct, Amsterdam, Netherlands

ACM ISBN 979-8-4007-1397-2/25/06
https://doi.org/10.1145/3707640.3731929

1 Introduction

Recent regulatory frameworks, such as the European Union's Digital Services Act (DSA) [13], highlight the increasing importance of managing the work process of auditing digital platforms at scale. Specifically, Article 34 of the DSA identifies four systemic risk categories and the obligation to assess online content through independent audits which are conducted by third-party auditing firms like KPMG or E&Y. Furthermore, the work to audit digital platforms under the DSA also concerns regulators in enforcement and researchers or NGOs as providers of evidence in the compliance process. Traditional audit and content moderation evaluation methods often rely heavily on human review processes, which can be labor-intensive and challenging to scale effectively. [10].

Advances in Large Language Models (LLMs) offer novel and expanded possibilities to address these challenges by automating the auditing process [1, 9, 17] and make it more scalable and cost-effective [8]. Yet, humans remain crucial in auditing content moderation quality and DSA compliance, particularly for complex [8, 11, 17] and interpretative tasks where LLMs may require significant fine-tuning and validation [5, 16]. This paper investigates the potential application of LLMs for auditing online content through automated classification and annotation techniques. To better understand the potential of these technologies in supporting the work of third-party auditors, researchers, and regulators, an interdisciplinary expert workshop with participants from law, social sciences/economics, and computer science was held at *TU Delft* in 2025.

Through collaborative discussions and interactive group work, participants drew on their diverse disciplinary perspectives to explore and critically assess model choices, parameter settings, and prompting techniques—highlighting how cross-domain expertise is essential for successfully navigating the technical, legal, economic, and societal complexities of auditing online platforms. This paper synthesizes the interdisciplinary discussions and evaluations, contributing to the ongoing debate about the potential and responsible integration of AI-based approaches into regulatory auditing practices to enable effective platform governance and meaningfully support the work of auditors.

Our results offer novel insights into the potential for LLM-supported DSA audits through structured workflows and iterative prompt design strategies in interdisciplinary teams to enhance accuracy, increase the audit scope and offer explanation and transparency of annotation decisions.

2 Related work

Recent studies recognize LLMs' utility in automating basic annotations [26, 29] and content moderation [14, 20], yet their limits in legal interpretation, historical content biases, cultural nuance, and hallucination resistance persist [3, 12, 19, 28, 34].

To address the inherent limitations of LLMs, structured workflows that combine human reviews and LLM-driven annotations have been proposed to improve reliability [24, 27, 33, 35]. Several authors suggest rigorous methodologies for annotation of content, proposing systematic prompt optimization, and fine-tuning to enhance annotation accuracy and reliability [8, 31, 32]. LLMs' ability to perform structured deductive coding methods using codebooks for streamlined annotation can differ in efficiency, and it depends on the content type (e.g., image or video) and structure of the material (e.g., comment to posts) to be analyzed [9, 11].

Certain LLM application scenarios demand precise alignment with legal standards — for example, auditing approaches under the DSA [4, 17, 25].

While LLMs can effectively assist human annotators, their standalone annotation capabilities for complex tasks remain inadequate. Therefore, cautious integration and validation strategies are warranted [5, 15, 16]. Our study expands this research by integrating interdisciplinary expert perspectives to evaluate the potential and feasibility of using LLMs for auditing political advertisements under the DSA's systemic risk framework.

3 Methodology

An expert workshop (N = 12) involving participants with interdisciplinary backgrounds and experience in collaborative, cross-disciplinary work was conducted at TU Delft in February 2025 (see Table 1 in the Appendix for an overview of participants' expertise).

The workshop comprised four main elements: (i) an introduction to the DSA and a methodology of LLM-supported content annotation for DSA audits; (ii) interdisciplinary breakout sessions for each team; (iii) a hands-on session refining LLM-supported audits in three teams (details below); and (iv) discussion and feedback.

Experts were divided into three interdisciplinary groups based on their primary and secondary topic preferences, expressed through the collaborative tool *Menti* [21]. Each group focused on a distinct key aspect: 1) The **Legal Team**: Examined how LLM annotations could support DSA compliance, evidence collection, and enforcement, emphasizing annotation's potential for systemic risk evaluation. 2) The **Social Science Team**: Focused on user perspective challenges, emphasizing accountability and transparency necessary for user trust in automated audits. 3) The **Tech Team**: Discussed the general potential of LLMs based on inherent weaknesses and identified strategies for improving LLM accuracy, emphasizing prompt engineering techniques to enhance overall reliability and measurement.

The results of the breakout and the hands-on sessions were documented on a Miro board [23] and subsequently analyzed qualitatively (see Fig. 1 in the Appendix for the documentation of the boards).

The workshop findings were analyzed using a structured qualitative content analysis [6, 7] to systematically derive insights from visual collaborative outputs.

First, we identified key themes of prominent textual elements (headings, bullet points, post-its, and highlighted texts), using color codes to distinguish the teams: Legal (white), Social Science (green), Tech (yellow) (see Fig. 3 in the Appendix). The thematic analysis of the Miro board outputs was conducted by the lead author, drawing on Braun & Clarke's reflexive thematic analysis framework [6]. The process began with a close reading of the visual data—including post-its, bullet points, and headings from the breakout and hands-on sessions. Initial codes were inductively derived from these visual elements, guided by the workshop structure (Legal, Social Science, Tech teams) and their respective color-coding (Legal – white, Social Science – green, Tech – yellow). The lead author and workshop organizer, whose background is in platform

governance and regulation, discussed iteratively the coding process with a co-author and co-organizer of the workshop, specializing in platform governance with a communication science background. These discussions helped to challenge assumptions, refine thematic categories, and ensure broader interdisciplinary validity.

In the next step, the extracted insights were systematically organized into thematic categories, reflecting regulatory compliance, annotation strategies, methodological challenges, and recommendations [7]. The visual clustering on the Miro board supported the transparency of this process, allowing for traceability from raw data to final themes. This process resulted in a set of higher-order themes reflecting shared concerns and recommendations across the interdisciplinary teams. An example coding trace is provided in the Appendix B to illustrate the development from raw data to thematic categories. In the 4 Results section, we synthesize and interpret the identified themes and insights into a coherent narrative across teams, contextualizing them within the existing literature on qualitative methods and annotation processes for legal compliance auditing.

4 Results

4.1 Breakout Sessions

The breakout sessions produced six thematic clusters: Technological Aspects, Organizational Aspects, Methodology and Quality Control, User Aspects, Platform-Related Aspects, and Regulation (see Fig. 3 in the Appendix).

Technological Aspects. Participants emphasized the value of structured prompts and efforts in explainability and control. Open-source models like LLaMA [22] and DeepSeek [18] were preferred "to ensure reproducibility." Setting the temperature parameter to zero was seen as key to reduce randomness and ensure consistency. Accessing LLMs via Application Programming Interfaces (APIs) allows to specify parameters and managing training/test data splits, contributing to greater methodological stability. Automatic prompt optimization using auxiliary models (e.g., Claude [2]) was considered helpful for improving prompt quality. Additionally, standardized formats (e.g. JSON) were suggested to facilitate processing and structure output.

Organizational Aspects. LLMs were seen as useful for updating codebooks and detecting rare categories. They can support the audit process as dialogic counterparts during the coding process, fostering reflection and documentation. One group proposed to "use the LLM to update the codebook itself," while another highlighted the potential to "annotate rare codes with a human to verify." However, assumptions about LLM neutrality were questioned—"no filtering in LLMs, or is there?"—underscoring the need for continued expert oversight.

Methodology and Quality Control. Participants called for repetitive testing and mandatory explanations in outputs. Concerns about automation and platform opacity were raised requiring outputs in automated processes to be reviewed by humans. Participants highlighted the need for model explanations to indicate how specific formulations or examples have impacted the output and enable effective prompt refinement. Further, making use of models' reasoning capacities may effectively reduce hallucinations [30]. Prompts should be tested repeatedly with varied phrasings to assess consistency. 'Few-shot prompting' and the use of an 'NA' option were debated due to potential downsides: the former can lead to overfitting on examples, while the latter may cause models to overuse 'NA' when facing even slight uncertainty.

User Aspects. Participants debated the limits of platform user empowerment through DSA-introduced rights—both as content creators and as voluntary contributors to audits via content reporting. They cautioned against excessive responsibilization, questioning, for example, "Why would a user have to explain that?" They emphasized that "being a skilled auditor is something separate," highlighting the distinction between professional auditors and everyday users. As such, they stressed that user-facing tools must include clear explanations and set appropriate expectations, to avoid losing valuable insights that users can offer within the audit scope.

Platform-Related Aspects. Discussions focused on inconsistent enforcement and lack of transparency. One participant asked whether platforms are "even enforcing" their own rules. Participants agreed that LLMs could help reveal enforcement gaps, if embedded in accountable processes.

Regulations. Finally, participants discussed the challenge of translating DSA requirements into LLM prompts. Structured legal definitions were seen as helpful, but overly technical language risked confusing the model. The consensus favored simplified legal guidance co-developed with domain experts.

4.2 Hands-On sessions: Evaluating prompt structures across teams

In the hands-on sessions, interdisciplinary teams collaboratively designed and tested prompt structures using multiple LLMs across diverse content formats documented in the Miro board (see Fig. 4 in the Appendix). Contributions clustered around: *Structure*, *Prompts*, *Testing*, *Legal Input*, and *Limitation of Scope*.

Structure. Participants emphasized the value of hierarchical prompts, where contextual framing, role-setting, and constrained output logic were clearly layered. JSON formatting and codebook references helped clarify classification schemes. As one team noted, inserting the image "as a comparison with text" helped validate shared context and reduce ambiguity in multimodal classification. However, others noted that model performance generally improved when images and especially videos were translated into descriptive text.

Prompts. Prompt construction varied across teams but typically included a mix of structuring the classification process for LLM use and setting a concrete context for categories. Common examples included: *"Please extract the Codenames and Definitions of the (XXX) Categories from the Codebook."* Others preferred multi-step prompts, beginning with legal or domain-specific framing and ending with specific instructions (e.g., *"Given the Codebook, identify a category from the image."*). Participants warned against overfitting to specific examples, emphasizing the need to generalize without reducing clarity.

Testing. Testing strategies involved small-scale experiments on representative instances. Incremental steps such as "test e.g., D1 first, then D2" were used to control for failure modes and isolate confounds. Testing across instances was seen as key, as missing factors may distort results. Prompts that explicitly asked for explanations

("demand explanations") were found to improve transparency and help identify weak reasoning.

Legal Input. Legal grounding was widely seen as essential for improving annotation relevance. Participants found that "legal definitions from the law work well," but flagged risks when using rare or overly complex legal terms. A preference emerged for using legal commentaries as contextual examples and allowing for "open definitions of a legal concept or several." Teams proposed that legal clarity in prompts supported both model accuracy and downstream audit explainability.

Limitation of Scope. Across all teams, prompt scoping was treated as a design strategy. Limiting prompt complexity by splitting up codebooks or restricting the prompt to a few categories increased classification precision. One participant noted that the "limitation of the knowledge base works well," supporting the idea that reduced scope leads to fewer hallucinations and better alignment. Participants advised against combining multiple tasks in a single prompt, favoring instead modularization and multi-prompt strategies.

5 Discussion: Contributions to Professional Practice and Platform Oversight

This study explored how LLMs can support systematic risk audits under the DSA through an interdisciplinary expert workshop. This approach emphasizes that the task is not merely a technical challenge, but a multidisciplinary design problem that requires coordination across legal interpretation, model behavior, and user-facing transparency.

The use of the visual Miro board supported rapid iteration, interdisciplinary comparison, and collective reflection. The session revealed that prompt engineering is a cross-cutting design task—spanning technical, legal, and methodological concerns essential for responsible LLM use in audits. This study offers a multidimensional contribution to the operationalization of systemic risk audits under the DSA by exploring how prompt engineering and LLM-supported annotation workflows can support auditors and other practitioners in their auditing work. Prompt design presents a multidisciplinary challenge that requires balancing legal clarity, model stability, and user transparency. The iterative experimentation across teams demonstrated that carefully constructed, focused, and legally grounded prompts produce more reliable and auditable outputs to support the work of DSA auditors —an essential requirement for applying LLMs to systemic risk audits under the DSA.

The findings demonstrate how structured prompts—especially those using hierarchical formats, legal references, and scope-limited inputs—can have scaling-up potential of audit scope. Participants emphasized the importance of rationale generation ("Let the LLM explain why it made a particular decision") as quality control mechanisms. These approaches increase reliability and interpretability—two main challenges for LLM use in auditing and crucial for systemic risk assessments.

The hands-on sessions provided meaningful insights for future work developing replicable frameworks for using LLMs in annotation contexts. The use of iterative testing, modular prompt decomposition, and multimodal scaffolding reveals how prompt design choices affect both model performance and human-AI collaboration. Importantly, participants highlighted the need for interdisciplinary co-design across social, economic, policy, legal, and technical domains to avoid misalignment between regulatory intent and model interpretation. This suggests new avenues for empirical studies on measurements such as intercoder reliability in hybrid (i.e., human + LLM) annotation designs and setups.

The study demonstrates how LLM-supported workflows can be designed to scale content analysis for enforcement and evidence collection while also creating documentation for other stakeholders at scale.

Structured prompts—with embedded legal logic, scoped queries, and demand for justifications—may enable regulators in testing audits to automate large-scale parts of the work process.

Participants highlighted how prompt chains and modular codebook design help manage complexity, allowing for targeted testing across diverse content formats and DSA risk categories. These methods allow practitioners not only to scale but also to systematically identify edge cases and iterate on category definitions. The findings suggest that AI-assisted auditing can support ongoing oversight and comparison across platforms, and support practitioner's work on the empirical evaluation of platform compliance over time.

Overall, this study highlights that employing LLMs to augment auditing is a design challenge at the intersection of domains that opens novel opportunities to audit at scale in the age of AI-based content generation. Therefore, designing human and LLM-based auditing approaches for the DSA requires methodological rigor, legal clarity, and context specific technical finetuning to ensure that audits remain interpretable, enforceable, and meeting regulatory demands.

6 Conclusion

This paper explored and assessed the feasibility and potential of employing LLMs for supporting the work process of auditing content, specifically targeting systemic risk evaluation under the DSA. Through an interdisciplinary expert workshop and hands-on prompt engineering sessions, we offered novel insights into both the potential and limitations of integrating LLMs into regulatory auditing practices. Our findings highlight that while LLMs provide significant opportunities for scalable and efficient annotation, their practical application requires careful calibration, structured prompting, and rigorous legal and methodological validation. The interdisciplinary approach underscored critical dimensions including legal precision, user-centric transparency, and technological robustness necessary for responsible deployment. This research represents an essential first step toward operationalizing AI-driven auditing within regulatory frameworks, emphasizing the value of and need for collaborative, multidisciplinary and multi-stakeholder expertise and exchange. Future work should focus on building end exploring a broader range of datasets, refining annotation design and methodologies, and developing iterative co-design processes that involve experts from multiple domains to facilitate that LLM-supported audits reliably enhance the enforceability of regulations.

Acknowledgments

We would also like to thank the participants of the workshop for their valuable contributions: John Albert, Shreyan Biswas, Max van Drunen, Lukas Selling, and Savvas Zanetthou, and Tom Viering.

We are grateful for the support of the Small Grants Scheme by the ELS Academy, who funded the workshop and this research study. Furthermore, this publication was supported by the AI Futures Lab on Rights and Justice, which is part of the TU Delft AI Labs programme and by the Federal Ministry of Education and Research of Germany (BMBF) under grant No. 16DII131 (Weizenbaum-Institut für die vernetzte Gesellschaft – Das Deutsche Internet-Institut).

References

[1] Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D. Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2024. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science* 8, 1 (Dec. 2024), 17. doi:10.1007/s42001-024-00345-9

[2] Anthropic. 2025. Claude Language Model. https://claude.ai/

[3] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. 2024. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3689904.3694699

[4] Marco Aspromonte, Andrea Ferraris, Federico Galli, and Giuseppe Contissa. 2024. LLMs to the Rescue: Explaining DSA Statements of Reason with Platform's Terms of Services. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis (Eds.). Association for Computational Linguistics, Miami, FL, USA, 205–215. doi:10.18653/v1/2024.nllp-1.17

[5] Savita Bhat and Vasudeva Varma. 2023. Large Language Models As Annotators: A Preliminary Evaluation For Annotating Low-Resource Language Content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, Daniel Deutsch, Rotem Dror, Steffen Eger, Yang Gao, Christoph Leiter, Juri Opitz, and Andreas Rücklé (Eds.). Association for Computational Linguistics, Bali, Indonesia, 100–107. doi:10.18653/v1/2023.eval4nlp-1.8

[6] Virginia Braun and Victoria Clarke. 2021. Conceptual and Design Thinking for Thematic Analysis. *Qualitative Psychology* 9 (May 2021), 3—-26. doi:10.1037/qup0000196

[7] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (July 2021), 328–352. doi:10.1080/14780887.2020.1769238

[8] Xiang Cheng, Raveesh Mayya, and João Sedoc. 2025. From Human Annotation to LLMs: SILICON Annotation Workflow for Management Research. doi:10.48550/arXiv.2412.14461 arXiv:2412.14461 [cs].

[9] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. doi:10.48550/arXiv.2306.14924 arXiv:2306.14924 [cs].

[10] Zachary Cooper, William Lehr, and Volker Stocker. 2024. The New Age: Legal & Economic Challenges to Copyright and Creative Economies in the Era of Generative AI. doi:10.2139/ssrn.5022340

[11] Edgar Dubourg, Valentin Thouzeau, and Nicolas Baumard. 2024. A step-by-step method for cultural annotation by LLMs. *Frontiers in Artificial Intelligence* 7 (May 2024), 1–10. doi:10.3389/frai.2024.1365508 Publisher: Frontiers.

[12] Shiran Dudy, Thulasi Tholeti, Resmi Ramachandranpillai, Muhammad Ali, Toby Jia-Jun Li, and Ricardo Baeza-Yates. 2025. Unequal Opportunities: Examining the Bias in Geographical Recommendations by Large Language Models. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1499–1516. doi:10.1145/3708359.3712111

[13] European Commission. 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) (Text with EEA relevance). http://data.europa.eu/eli/reg/2022/2065/oj/eng Legislative Body: EP, CONSIL.

[14] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2024. Integrating Content Moderation Systems with Large Language Models. *ACM Trans. Web* 0 (Oct. 2024), 1 – 20. doi:10.1145/3700789 Just Accepted.

[15] Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès, and Dan Jurafsky. 2025. Can Unconfident LLM Annotations Be Used for Confident Conclusions? doi:10.48550/arXiv.2408.15204 arXiv:2408.15204 [cs].

[16] Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large Language Models Are Effective Human Annotation Assistants, But Not Good Independent Annotators. doi:10.48550/arXiv.2503.06778 arXiv:2503.06778 [cs].

[17] Andrew Halterman and Katherine A. Keith. 2025. Codebook LLMs: Evaluating LLMs as Measurement Tools for Political Science Concepts. doi:10.48550/arXiv.2407.10747 arXiv:2407.10747 [cs].

[18] Hangzhou DeepSeek Artificial Intelligence Basic Technology Research Co., Ltd. 2025. DeepSeek. https://www.deepseek.com/

[19] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[20] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3613905.3650828

[21] Mentimeter. 2025. Menti Online Voting. https://www.menti.com/

[22] Meta AI. 2025. Large Language Model Meta AI. https://www.llama.com/

[23] Miro. 2025. Miro Collaboration Platform. https://miro.com/

[24] Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. doi:10.48550/arXiv.2406.06369 arXiv:2406.06369 [cs].

[25] Jakob Mökander, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 2024. Auditing large language models: a three-layered approach. *AI and Ethics* 4, 4 (Nov. 2024), 1085–1115. doi:10.1007/s43681-023-00289-2

[26] Arbi Haza Nasution and Aytuğ Onan. 2024. ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access* 12 (2024), 71876–71900. doi:10.1109/ACCESS.2024.3402809 Conference Name: IEEE Access.

[27] 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*. ACM, Montréal QC Canada, 913–926. doi:10.1145/3600211.3604712

[28] Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2541–2573. doi:10.18653/v1/2023.emnlp-main.155

[29] Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large Language Models for Data Annotation and Synthesis: A Survey. doi:10.48550/arXiv.2402.13446 arXiv:2402.13446 [cs].

[30] Stephen L Thaler. 1995. "Virtual input" phenomena within the death of a simple pattern associator. *Neural Networks* 8, 1 (1995), 55–65.

[31] Petter Törnberg. 2023. How to use LLMs for Text Analysis. doi:10.48550/arXiv.2307.13106 arXiv:2307.13106 [cs].

[32] Petter Törnberg. 2024. Best Practices for Text Annotation with Large Language Models. doi:10.48550/arXiv.2402.05129 arXiv:2402.05129 [cs].

[33] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–21. doi:10.1145/3613904.3641960

[34] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao 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. arXiv:2309.01219 [cs.CL] https://arxiv.org/abs/2309.01219

[35] Chengshuai Zhao, Zhen Tan, Chau-Wai Wong, Xinyan Zhao, Tianlong Chen, and Huan Liu. 2025. SCALE: Towards Collaborative Content Analysis in Social Science with Large Language Model Agents and Human Intervention. arXiv:2502.10937 [cs.AI] https://arxiv.org/abs/2502.10937

A Information About Workshop Participants

Table 1: Participant Expertise and Years of Experience.

#	Area of Expertise	Experience
1	HCI	<2 years
2	Law, Information Systems, Journalism	2–5 years
3	Security, Privacy, Measurements	>8 years
4	Consumer IoT behaviours, Security Awareness, Security Economics	>8 years
5	Computational Social Science, Hate Speech, Shitposting	5–8 years
6	Transformers, Machine Learning	5–8 years
7	Data Access, Research Engineering	2–5 years
8	HCI, ML	2–5 years
9	HCI, UX, Qualitative Coding	>8 years
10	Digital Infrastructures, Platforms, and Services; Competition, Regulation, Innovation	>8 years
11	Platform Governance, Disinformation, DSA	2–5 years
12	Platform Regulation, Blogging	>8 years

B Thematic Analysis Process and Coding Trace

B.1 Thematic Analysis Process

The thematic analysis followed Braun & Clarke's reflexive framework [6, 7], using an inductive approach to derive themes from the Miro board outputs.

The analysis comprised the following steps:

(1) **Familiarization with Data**: The lead author reviewed the Miro board content, focusing on the color-coded outputs from the three interdisciplinary teams: Legal (white), Social Science (green), and Tech (yellow).
(2) **Generating Initial Codes**: Codes were created inductively, capturing key concepts and recurring patterns within each team's contributions.
(3) **Collating Codes into Themes**: Codes were clustered into higher-order themes that spanned across teams, reflecting shared or divergent insights.
(4) **Reviewing Themes**: The initial themes were iteratively refined through discussions with a co-author specializing in platform governance and data regulation. These discussions challenged assumptions and ensured alignment with interdisciplinary perspectives.
(5) **Defining and Naming Themes**: Final themes were labeled to reflect key concerns and actionable insights relevant to LLM-supported auditing under the DSA.
(6) **Producing the Report**: Themes were integrated into the *Results* section and contextualized with relevant literature. Traceability from raw data to final themes is demonstrated in Table 2.

B.2 Detailed Coding Trace

Table 2: Sample of Coding Traces from Miro Board to Final Themes

Raw Miro Data (Post-it / Note)	Initial Code	Final Theme	Team
Use the LLM to update the codebook itself	Codebook Adaptation	Organizational Aspects	Tech
Annotate rare codes with a human to verify	Human-in-the-loop Verification	Organizational Aspects	Legal
Demand explanations in outputs	Explanation Requirement	Methodology & Quality Control	Social Sci.
Prompts should be tested repeatedly with varied phrasings	Prompt Testing Strategies	Methodology & Quality Control	Tech
Users shouldn't be responsible for explaining everything	User Accountability	User Aspects	Social Sci.
Being a skilled auditor is something separate	Professionalization of Auditing	User Aspects	Social Sci.
Are platforms even enforcing their own rules?	Enforcement Gaps	Platform-Related Aspects	Legal
Simplified legal guidance co-developed with experts	Legal Simplification	Regulations	Legal
Open-source models preferred to ensure reproducibility	Model Reproducibility	Technological Aspects	Tech
Set temperature to zero to reduce randomness	Parameter Tuning	Technological Aspects	Tech
Legal definitions from law work well but too complex terms can fail	Legal Framing of Prompts	Legal Input	Legal
Prompt scoping helps reduce hallucinations	Scope Limitation	Limitation of Scope	Tech

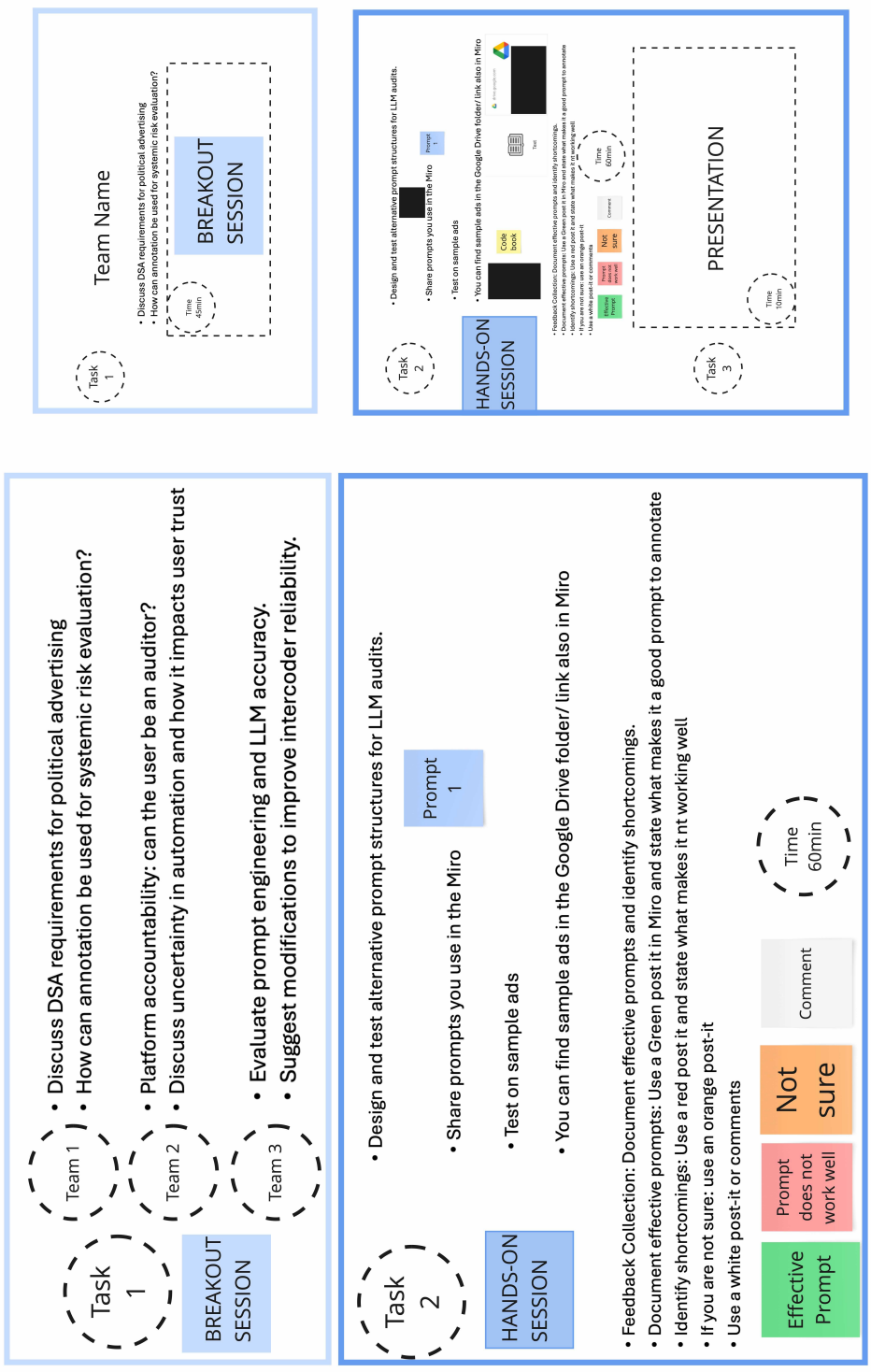


Figure 1: Zoomed-in visual documentation of the Breakout and Hands-On Sessions via Miro board, showing the structure of tasks and layout from which thematic analysis was conducted. The board aggregates key ideas and discussion points on leveraging LLMs for systemic risk evaluation under the DSA.

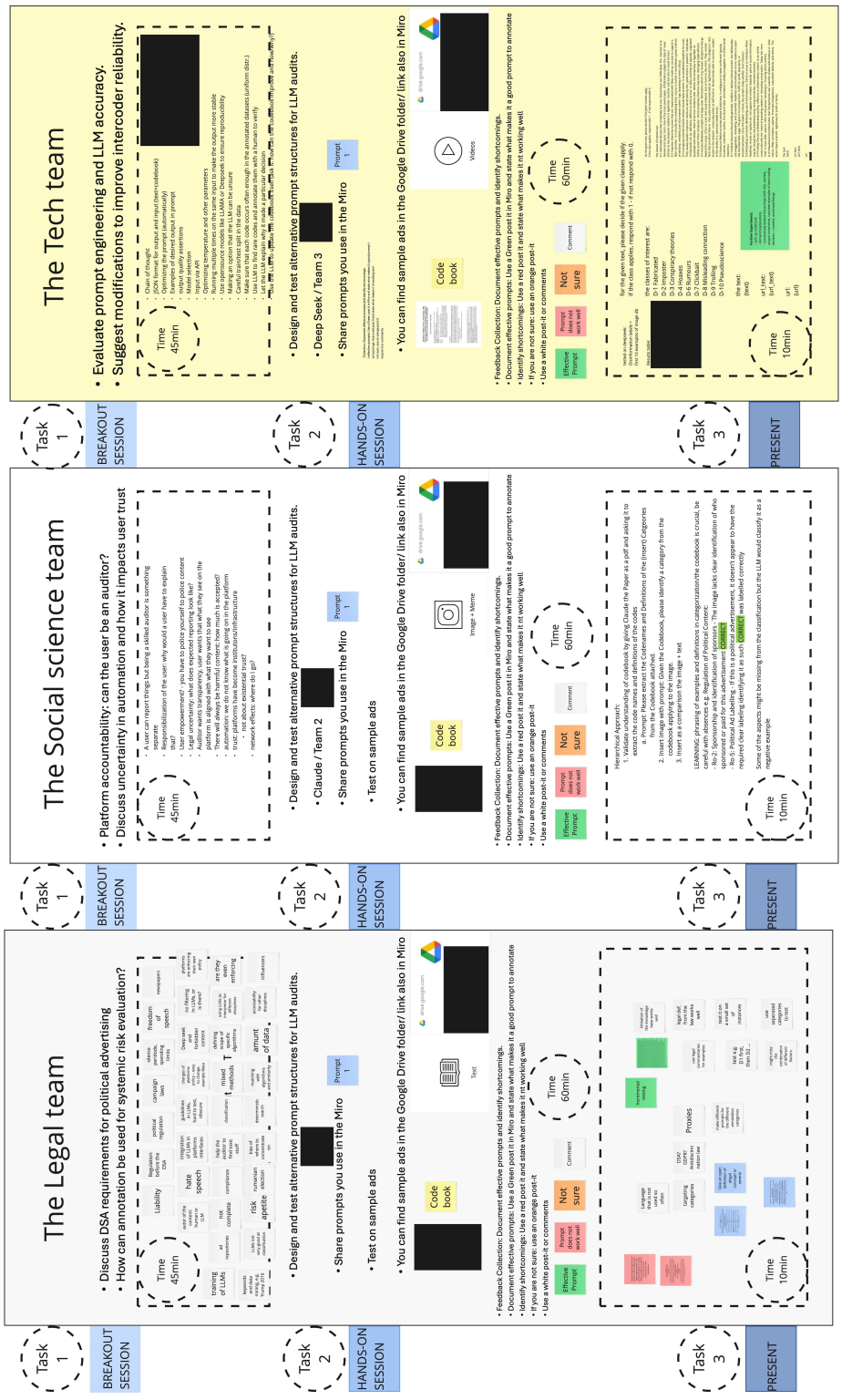


Figure 2: Zoomed-out overview of all Breakout and Hands-On Sessions Miro boards, visualizing inputs from the interdisciplinary expert workshop on LLM-based systemic risk evaluation under the DSA.

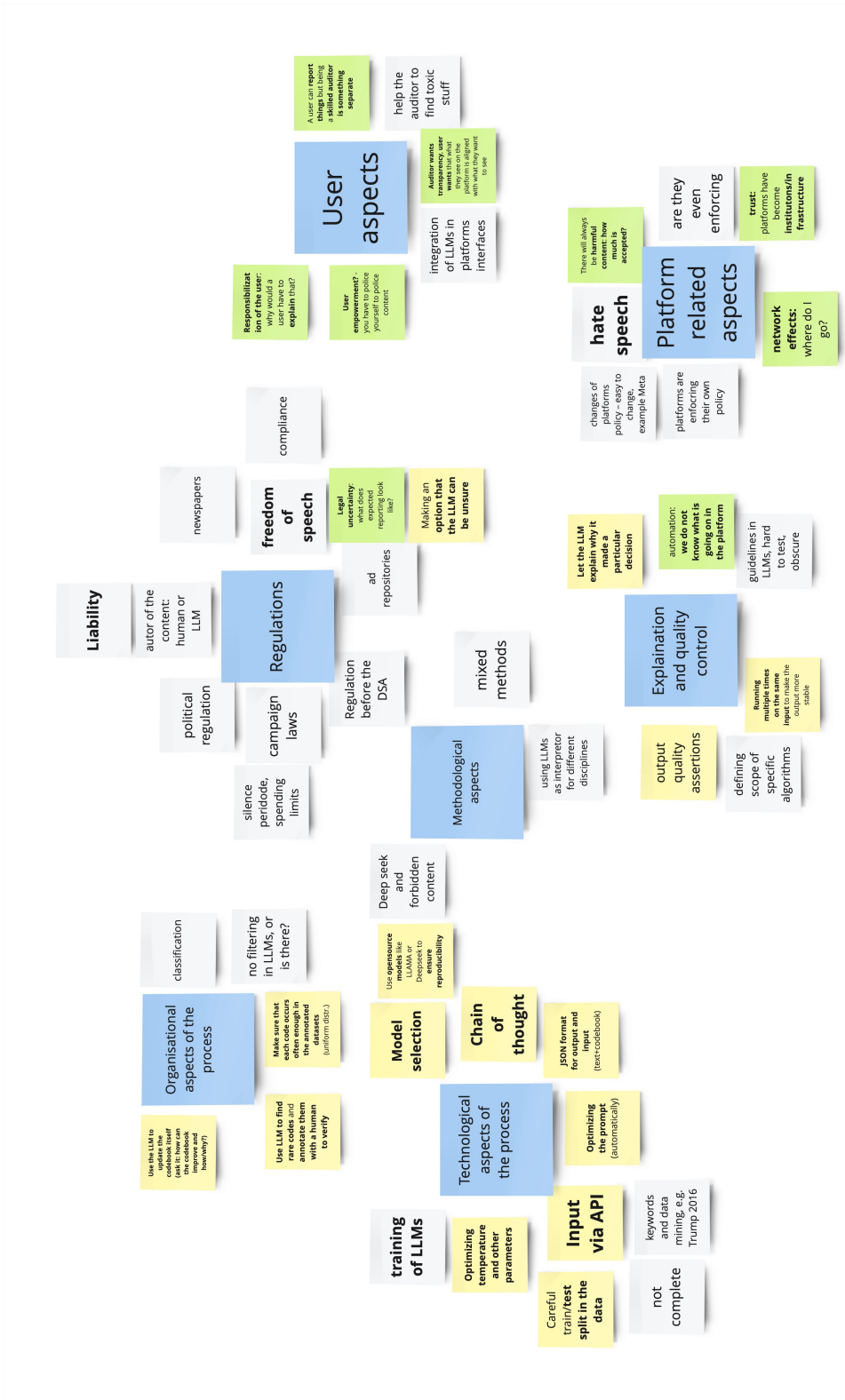


Figure 3: Clustered and color-coded visualization of the Breakout Session discussions, as captured on the Miro board. These inputs were synthesized through thematic analysis, with key insights detailed in the text.

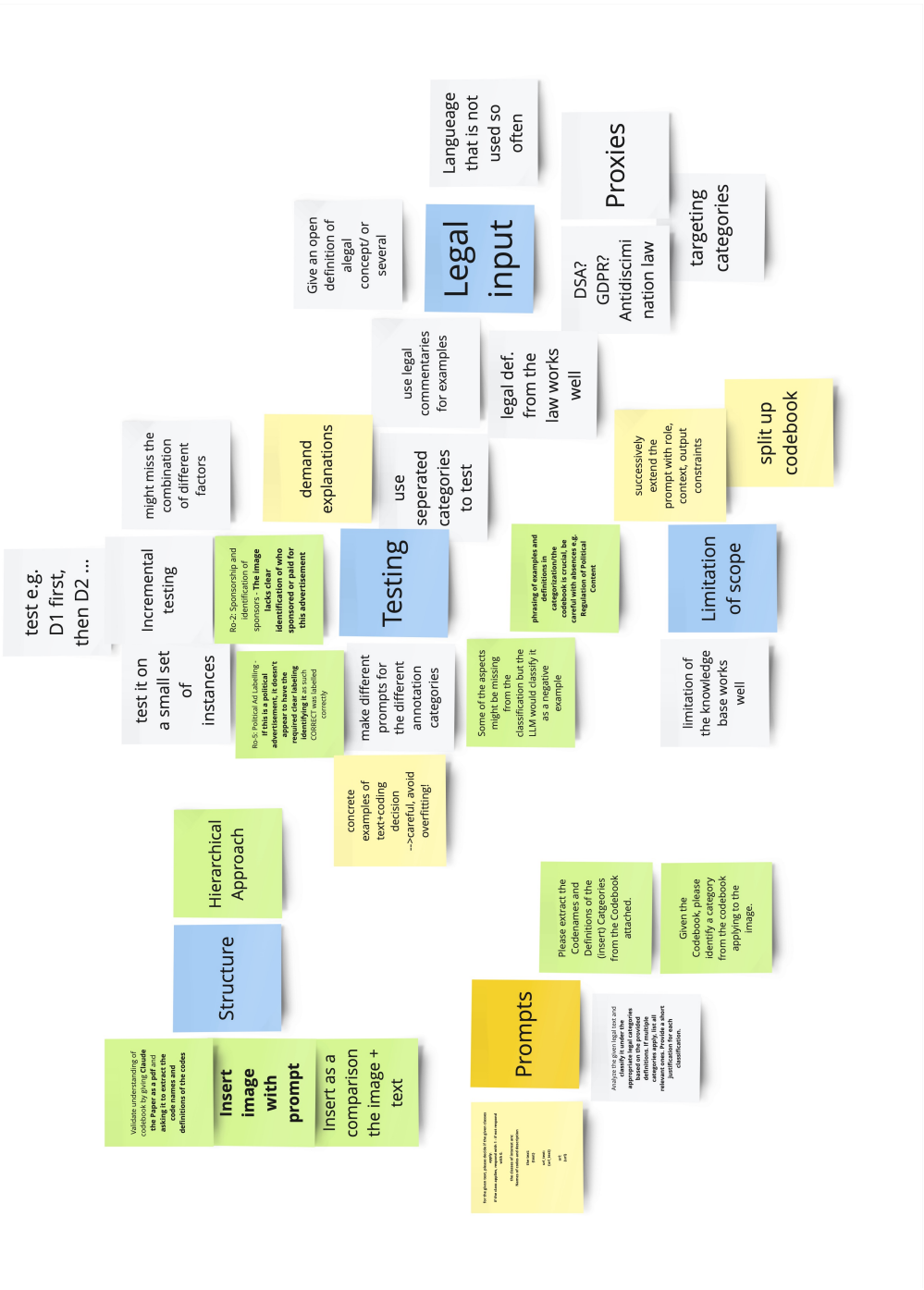
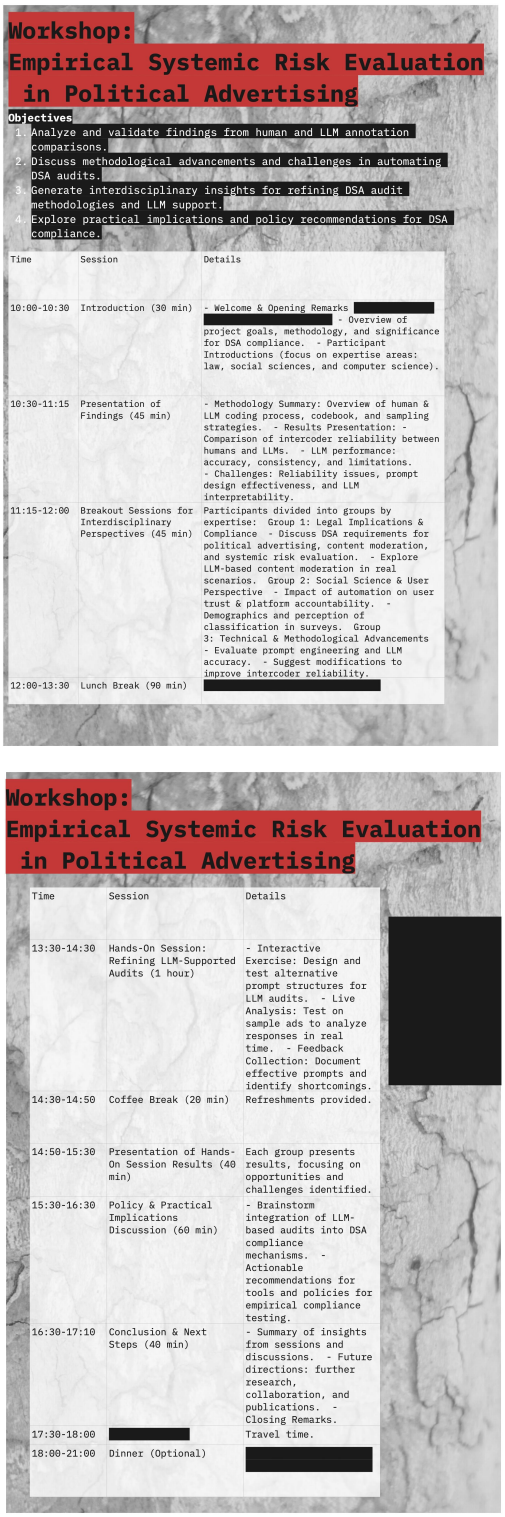


Figure 4: Clustered and color-coded visualization of the Hands-On Session discussions, as captured on the Miro board. These inputs were synthesized through thematic analysis, with key insights detailed in the text.



Workshop: Empirical Systemic Risk Evaluation in Political Advertising

Objectives

1. Analyze and validate findings from human and LLM annotation comparisons.
2. Discuss methodological advancements and challenges in automating DSA audits.
3. Generate interdisciplinary insights for refining DSA audit methodologies and LLM support.
4. Explore practical implications and policy recommendations for DSA compliance.

Time	Session	Details
10:00-10:30	Introduction (30 min)	- Welcome & Opening Remarks [redacted] - Overview of project goals, methodology, and significance for DSA compliance. - Participant Introductions (focus on expertise areas: law, social sciences, and computer science).
10:30-11:15	Presentation of Findings (45 min)	- Methodology Summary: Overview of human & LLM coding process, codebook, and sampling strategies. - Results Presentation: - Comparison of intercoder reliability between humans and LLMs. - LLM performance: accuracy, consistency, and limitations. - Challenges: Reliability issues, prompt design effectiveness, and LLM interpretability.
11:15-12:00	Breakout Sessions for Interdisciplinary Perspectives (45 min)	Participants divided into groups by expertise: Group 1: Legal Implications & Compliance - Discuss DSA requirements for political advertising, content moderation, and systemic risk evaluation. - Explore LLM-based content moderation in real scenarios. Group 2: Social Science & User Perspective - Impact of automation on user trust & platform accountability. - Demographics and perception of classification in surveys. Group 3: Technical & Methodological Advancements - Evaluate prompt engineering and LLM accuracy. - Suggest modifications to improve intercoder reliability.
12:00-13:30	Lunch Break (90 min)	[redacted]

Workshop: Empirical Systemic Risk Evaluation in Political Advertising

Time	Session	Details
13:30-14:30	Hands-On Session: Refining LLM-Supported Audits (1 hour)	- Interactive Exercise: Design and test alternative prompt structures for LLM audits. - Live Analysis: Test on sample ads to analyze responses in real time. - Feedback Collection: Document effective prompts and identify shortcomings.
14:30-14:50	Coffee Break (20 min)	Refreshments provided.
14:50-15:30	Presentation of Hands-On Session Results (40 min)	Each group presents results, focusing on opportunities and challenges identified.
15:30-16:30	Policy & Practical Implications Discussion (60 min)	- Brainstorm integration of LLM-based audits into DSA compliance mechanisms. - Actionable recommendations for tools and policies for empirical compliance testing.
16:30-17:10	Conclusion & Next Steps (40 min)	- Summary of insights from sessions and discussions. - Future directions: further research, collaboration, and publications. - Closing Remarks.
17:30-18:00	[redacted]	Travel time.
18:00-21:00	Dinner (Optional)	[redacted]

Figure 5: Program for Expert Workshop.