Assessing Compliance for Generative AI Systems under the Digital Services Act

Marie-Therese Sekwenz[1], Rita Hermann-Gsenger[2], Shreyan Biswas[3] and Ujwal Gadiraju[4]

[1]Technology & Policy Management, Delft University of Technology, Delft, The Netherlands
[2]Weizenbaum Institute, Berlin, Germany
[3]Software Technology Department, Delft University of Technology, Delft, The Netherlands
[4]Software Technology Department, Delft University of Technology, Delft, The Netherlands

Abstract

Generative AI systems such as ChatGPT are increasingly embedded into online services that shape information access, content creation, and user decision-making at scale. This raises a timely governance question for European digital regulation: how should large language model (LLM) services be assessed under the Digital Services Act (DSA), and what would compliance auditing require if such services qualify as Very Large Online Platforms or Search Engines (VLOPs/VLOSEs)? While the DSA's oversight regime was designed around platform-era risks—such as user-generated content, recommender systems, and notice-and-action pipelines—LLM services express many compliance-relevant decisions through interactive safeguards, including refusals, safety warnings, and guardrail-driven output constraints. These features create new auditing challenges around reproducibility, transparency of enforcement logic, and the evaluation of model-generated content pathways. In this position paper, we analyse different forms of generative AI including ChatGPT-style services within DSA compliance structures and how generative AI reshapes assessment and, mitigation practices particularly in interaction with the AI Act. We outline a DSA-anchored testing perspective that maps core legal obligations to LLM-tailored audit considerations and structured prompt-based evaluation strategies. Finally, we clarify key intersections between the DSA and the AI Act, showing how service-level systemic risk assessment (Art. 34 DSA) and model-level governance obligations under the AI Act (e.g., Art. 3(66)) interact in practice. We argue that effective oversight of generative AI-infused services requires prompt- and risk-based systemic assessment and independent audit designs that draw on multiple compliance signals rather than solely traditional moderation metrics.

Keywords

Digital Services Act, Large Language Model, ChatGPT, VLOP, Artificial Intelligence Act

Large Language Model (LLM) services such as ChatGPT are rapidly becoming core infrastructures for information access [1, 2], content creation [3, 4, 5], user interaction online [6], and for executing everyday tasks [7]. As these systems scale and integrate into consumer-facing products and shape enterprise workflows [8], they increasingly function as large scale intermediaries that structure information exposure, user behaviour, and potential harm at population level [9, 10]. Regulators have begun to examine whether and how such services fall within existing European legal frameworks, particularly where generative AI features are embedded in large platforms or provide search-like functionality at scale. At the same time, major online platforms are integrating LLM-powered interfaces directly into core services, further blurring the boundary between AI systems and intermediary platforms.

The Digital Services Act (DSA) is the European Union's horizontal regulatory framework for online intermediary services, including online platforms and search engines [11].[1] It establishes due-diligence obligations aimed at identifying, assessing, and mitigating systemic risks, including risks to fundamental rights, civic discourse, and the dissemination of illegal content [12]. In parallel, the AI Act, adopted in 2024, introduces a distinct regulatory regime for AI systems, including general purpose AI models such as large language models [13, 14]. While the DSA governs service level risks arising from large scale online intermediaries, the AI Act focuses on system and model level governance, including risk management,

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
[†]These authors contributed equally.
M.T.Sekwenz@tudelft.nl (M. Sekwenz); rita.gsenger@weizenbaum-institut.de (R. Hermann-Gsenger); S.Biswas@tudelft.nl (S. Biswas); U.K.Gadiraju@tudelft.nl (U. Gadiraju)

[1]In this work we concentrate on the DSA as the regulatory concept fully in force, while also foreshadowing the main concepts of the AI Act, like human oversight (Art. 14 AI Act) for high risk systems.

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-06-27

transparency, and human oversight obligations. These regulations raise a pressing doctrinal and methodological challenge for EU digital governance: *How should LLM-based services be assessed under the DSA's systemic risk and audit framework, particularly where they qualify as VLOPs/VLOSEs, and how does this assessment intersect with AI Act obligations?* A central innovation of the DSA is the obligation for designated Very Large Online Platforms and Very Large Online Search Engines (VLOPs/VLOSEs) to identify, assess, and document systemic risks (Art. 34 DSA), supported by periodic independent audits (Art. 37 DSA) [15]. These provisions institutionalize structured oversight at the service level, requiring platforms that reach significant scale thresholds (Art. 33 DSA) to evaluate how their design, functionalities, and governance processes may generate societal risks.

Applying this framework to LLM based services is non trivial. The DSA's strongest mechanisms were developed in response to platform era risks primarily associated with user generated content and recommender systems. In those settings, platforms primarily act as intermediaries: they host, rank or remove content created by users and audits examine how moderation decisions are taken and documented.

LLM services complicate this intermediary model. Here, the service itself generates content in response to user prompts, and compliance-relevant decisions may be embedded directly within the generation process. As a result, the object of assessment under the DSA shifts. Risks may arise not only from how user content is handled, but also from how the system produces, constrains, and presents information.[2] Auditing must therefore examine generative processes and embedded enforcement mechanisms through which exposure is structured, rather than focusing exclusively on moderation decisions applied to user-generated content.

Beyond this structural shift, generative AI introduces additional challenges for regulatory oversight. First, attribution and responsibility are reconfigured when the service itself produces potentially harmful outputs. Second, conversational interactions are often ephemeral and context-dependent, complicating traceability and reproducible sampling for systemic risk assessments. Third, enforcement decisions are embedded within output behaviour, making it harder to distinguish between moderation, system design constraints, and ordinary model responses.

These features generate interpretive and methodological challenges for both systemic risk assessment (Art. 34 DSA) and independent audits (Art. 37 DSA), particularly where DSA obligations intersect with AI Act requirements concerning risk management (Art. 9 AI Act) and oversight (Art. 14 AI Act).

1. ChatGPT as VLOP/VLOSE: A Qualification Question—Why Designation Status Matters for Oversight

The DSA's most stringent obligations apply to services designated as VLOPs or VLOSEs (Art. 33 DSA). Designation triggers requirements to conduct annual systemic risk assessments (Art. 34 DSA), implement proportionate mitigation measures (Art. 35 DSA), [3] and undergo external independent audits conducting independent auditing firms (Art. 37 DSA). In other words, designation determines whether a provider is subject to the DSA's strongest accountability and oversight mechanisms.

For LLM-based services, qualification is not merely a formal threshold question but a structural governance issue. The DSA applies to online intermediary services that meet quantitative criteria (notably the threshold of 45 million average monthly active recipients in the EU according to Art. 33 DSA). The regulatory challenge arises in determining, whether standalone LLM services qualify as online platforms or search engines under the DSA, and how LLM functionalities embedded within already designated VLOPs/VLOSEs alter the scope of systemic risk assessment [16]. The following examples demonstrate how such LLM systems are integrated:[4]

[2] For example the generation of nudes or bikini depictions of images through Grok, or the usage of Google AI mode in generating election advice.

[3] Risk mitigation and risk assessment are a connected part of DSA compliance that goes hand in hand. First the risks must be identified, assessed, and then mitigated. These steps must be documented in the risk assessment reports issued by VLOPs and VLOSEs.

[4] These examples all fall (soon) under the DSA as a VLOP/VLOSE. All of them would also qualify as an AI system (Art. 3(66) AI

- **Search engine augmentation (e.g., Bing AI Search):** Where a designated VLOSE incorporates generative AI into core search functionality [17], the LLM component becomes part of the service's systemic risk profile and must be reflected in Art. 34 risk assessments and Art. 37 audits.
- **Platform-embedded chatbots (e.g., X and Grok):** Where a designated VLOP integrates an LLM as a feature, generative outputs are produced within a regulated platform environment. The chatbot does not create a separate regulated entity but modifies the risk surface of the existing service.
- **Standalone LLM services (e.g., ChatGPT):** Where an LLM service provides large-scale interactive information access and potentially search-like functionality, questions arise as to whether it meets the definitional and quantitative criteria for designation under the DSA [18].

Conclusion: Importantly, designation regulates the service through which an LLM is deployed, not the model as such. As LLM-based services and features become embedded in large-scale digital infrastructures, DSA oversight must grapple with how to apply systemic risk assessment and audit duties to generative outputs, interaction flows (like the integration of notice and action mechanisms according to Art. 16 DSA to refer platforms to AI generated content through UIs), and AI labelling (Art. 34(2)(a) DSA, Art. 40(b) AI Act).[5] Accordingly, compliance questions concerning generative AI services cut across multiple DSA provisions, including transparency of ToS (Art. 14), notice-and-action mechanisms (Art. 16), complaint-handling procedures (Arts. 20–22), statements of reasons (Art. 17), and systemic risk assessment and audit obligations (Arts. 34 and 37). Consequently, auditing shifts beyond conventional notice-and-takedown and content moderation (e.g. for detecting spam or automated attacks in the context of AI content), towards evaluating the socio-technical conditions under which AI-systems generate, amplify, and structure information exposure at scale. The central methodological question is how these provisions should be operationalised when moderation logic is embedded within generative systems rather than applied externally to static content objects.

2. Why ChatGPT Changes the DSA Compliance Problem

The DSA operationalises due diligence for hosting services, platforms, and search engines, which must address illegal content and ToS violations while safeguarding fundamental rights. In platform era settings, compliance relevant decisions typically concern the handling of user generated content: removal, labelling, ranking adjustments, or account restrictions. LLM-based services complicate this structure. Here, compliance-relevant decisions are frequently expressed through *interactive constraints*: refusals, partial compliance, safety redirections, and capability limitations. Rather than reacting to user-created content, the service generates outputs and simultaneously constrains them. This alters the compliance object under the DSA from moderation of hosted content to governance of generative output pathways. This shift introduces three interrelated challenges for systemic risk assessment (Art. 34 DSA) and independent audits (Art. 37 DSA):

1. **Reconfigured compliance object.** Risks may arise not only from how user content is handled, but from how the system produces, frames, and constrains information. Audits must therefore evaluate generative behaviour in addition to conventional moderation workflows.
2. **Methodological implications for risk assessment.** Conversational interactions are context dependent and may be ephemeral, complicating traceability, reproducibility, and representative sampling under Art. 34 DSA.[6]

Act), or even as a General Purpose System or model, triggering high risk classification under the AI AI with a more stringent rule corset, and can even lead up to a systemic risk classification under the AI Act.

[5] According to the AI Act's view on labelling for 'deepfake' content should consider "aspects such as graphics, position and timing, drawing on scientific research on the effectiveness of labels."

[6] **Arts. 34 and 37 (systemic risk assessment and independent audits):** Where do LLMs demand different auditing structures compared to search engines and social media platforms, and how does the methodology change according to the Delegated Act to identify, assess, and mitigate risks?

3. **Embedded enforcement and interface design.** Refusals, warnings, and labelling practices function as de facto enforcement mechanisms. Their consistency, transparency, and alignment with communicated ToS obligations become relevant for Arts. 14, 16, and 17 DSA, as well as for complaint-handling, out-of-court-dispute-settlement, and trusted flagger procedures (Arts. 20–22).[7]

2.1. ChatGPT—A Generative AI System under the AI Act

ChatGPT constitutes a *general-purpose AI* (GPAI) system within the meaning of Art. 3(66) AI Act, as it can be used across a wide range of tasks and contexts, including downstream integration into other products and services. Where such a system is trained with large-scale compute and exhibits broad capabilities, it may also fall within the AI Act's category of *GPAI models with systemic risk* (Art. 3(65) AI Act), which triggers additional provider obligations (e.g. Art. 50).

2.2. Main Obligations under the AI Act

The AI Act imposes a layered set of duties depending on whether the system is (i) a GPAI system, (ii) a GPAI model with systemic risk, or (iii) deployed as part of a *high-risk* AI system (Annex III AI Act). Key obligations include technical documentation and transparency duties, information provision to downstream deployers, and measures to enable *human* oversight and accountability across the AI system lifecycle, including risk management, data governance, cyber security, fundamental rights impact assessment.

2.3. Systemic Risk under the AI Act Is Not the DSA's Systemic Risk Assessment

Although both regimes use the term *systemic risk*, the AI Act's systemic-risk concept is not equivalent to the DSA's systemic risk assessment. Under the DSA, systemic risk assessments (Art. 34, 37 DSA) focus on platform-level, society-facing risks that arise from the design and functioning of services at scale (e.g., illegal content, fundamental rights impacts, and risks to civic discourse), and are tied to designation as a VLOP/VLOSE. In contrast, the AI Act's systemic-risk regime targets model-level and capability-related risks associated with GPAI models (Art. 3(63) AI Act), emphasising technical governance, and safety. For ChatGPT compliance assessment, this creates an intersection: DSA duties structure oversight of the *service and its societal effects*, while AI Act duties structure governance of the *model and its deployment lifecycle*.

3. Position

We argue that DSA oversight for LLM services cannot rely exclusively on audit practices developed for social media moderation – but must instead anticipate the regulatory overlap between the AI Act and the DSA. Where moderation logic is embedded within generative systems, compliance assessment must examine how outputs are produced, constrained, and presented in response to user prompts.

[7] **Art. 14 (ToS):** Referring to rules for dealing with genAI content in place, and are enforcement outcomes consistent with communicated rules? **Art. 16 (notice-and-action):** Can users submit actionable (illegal content) notices, and is the design easy to access and user-friendly? **Art. 20 (internal complaint-handling):** Are complaint channels supervised by qualified staff rather than solely automated? (E.g. if a user's content is wrongly flagged as AI generated)? **Art. 21 (out-of-court-dispute-settlement):** Are complaint channels to out-of-court dispute settlement bodies possible (e.g., can a user complain about a platform decision not to take down "not similar enough" deepfakes of a person)? **Art. 22 (trusted flaggers):** Can trusted flaggers register, be prioritised, and is their work registered in reporting genAI content through dedicated UIs as external human oversight stakeholders? **Art. 17 (statement of reasons):** When restrictions are imposed (e.g., disabling access, account action), are reasons clear and specific—including when automated means are used (AI systems); is genAI content identified as the database category "synthetic data"; and how would reporting with the database be operationalised in chatbot functions (including whether appearances of guardrails are taken into account as a moderation measure)?

We therefore propose *prompt-based compliance testing* [8] as a complementary methodological lens: structured and reproducible evaluation of moderation-like decisions expressed through interactive safeguards (e.g., refusals, warnings, guardrails, labels) and governance workflows (e.g., account restrictions, explanations, illegal content reporting). For example, structured prompt sets could be used to assess refusal consistency across variations of illegal content requests or to evaluate how election-related queries are framed under different interaction contexts.

Such testing aligns with sampling methodologies recognised under the Delegated Act for supporting Art. 34 and Art. 37 DSA assessments [19]. It should further account for time-sensitive dynamics (e.g., virality effects [20]) and context-dependent harms such as disinformation [21].

3.1. Contributions

We make three contributions understanding DSA assessment as a form of human oversight and connected to the AI Act in the context of AI systems and AI generated content on platforms. First, we propose an overview to analyse whether, and in which parts ChatGPT-style services may fall within the DSA's regulatory categories, distinguishing between standalone LLM services and LLM functionalities embedded within designated VLOPs/VLOSEs. Second, we outline a DSA-anchored testing perspective that maps concrete legal obligations to LLM-tailored audit procedures, building on existing systemic risk reporting practices. Third, we outline key intersections between the AI Act and the DSA demonstrating how model-level and service-level overlapping obligations jointly structure meaningful human oversight in the domain of risk assessment under EU digital regulation.

3.2. Open Research Directions

We propose three research directions suitable for an interdisciplinary workshop agenda: First, **qualification and scope questions** concerning LLM services under EU regulation require further clarification, including how standalone generative systems and embedded LLM features should be classified under the DSA and how designation status interacts with AI Act obligations.

Second, **methodological adaptations for systemic risk assessment** are needed for generative services. Unlike traditional platform auditing, which evaluates moderation decisions applied to user-generated content moderation strategies, LLM-based services represent human–AI interaction. Oversight may therefore require structured sampling of interactional outputs, red-teaming, reproducibility protocols, and behavioural evaluation methods [5, 22] that assess how users and systems co-produce content under varying prompts and interface conditions.

Third, **overlapping regulatory obligations and interface design** warrant closer examination. The interaction between DSA provisions (e.g., notice-and-action, complaint handling, statements of reasons) and AI Act requirements (e.g., transparency, labelling, and human oversight) raises questions about how reporting mechanisms, guardrails, and explanations should be operationalised within conversational user interfaces.

Together, they foreground the practical and methodological frictions that emerge when DSA auditing is extended to LLM-based services and other generative AI features, that also touch the scope of the AI Act. These directions align closely with the core concern of the AI CHAOS workshop: *poorly designed oversight* in auditing and assessment tasks around AI-systems and content can itself generate auditing risk.

[8] By *prompt-based* testing we refer to tests that include red-teaming strategies, document guardrail effectiveness, use different personas, temperatures, or documenting hallucination counts for election information.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT (OpenAI) to support language editing, including grammar and spelling checks, sentence polishing, and the rephrasing of author-written text for clarity and conciseness. The tool was not used to generate scientific insights, conduct the analysis, interpret the findings, draw conclusions, or formulate recommendations. All AI-assisted text was critically reviewed, verified, and edited by the authors, who take full responsibility for the content of this publication.

References

[1] R. W. White, C. Shah, Information Access in the Era of Generative AI, Springer, 2025.

[2] A. Rieger, T. Draws, N. Mattis, D. Maxwell, D. Elsweiler, U. Gadiraju, D. McKay, A. Bozzon, M. S. Pera, Responsible opinion formation on debated topics in web search, in: European Conference on Information Retrieval, Springer, 2024, pp. 437–465.

[3] Z. Epstein, A. Hertzmann, I. of Human Creativity, M. Akten, H. Farid, J. Fjeld, M. R. Frank, M. Groh, L. Herman, N. Leach, et al., Art and the science of generative ai, Science 380 (2023) 1110–1111.

[4] Y. Hua, S. Niu, J. Cai, L. B. Chilton, H. Heuer, D. Y. Wohn, Generative ai in user-generated content, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–7.

[5] S. Biswas, A. Erlei, U. Gadiraju, Mind the gap! choice independence in using multilingual llms for persuasive co-writing tasks in different languages, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–20.

[6] J. D. Weisz, J. He, M. Muller, G. Hoefer, R. Miles, W. Geyer, Design principles for generative AI applications, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–22.

[7] G. He, G. Demartini, U. Gadiraju, Plan-then-execute: An empirical study of user trust and team performance when using llm agents as a daily assistant, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–22.

[8] U. Gadiraju, A. Balayn, The enterprising and elusive prospects of human-ai collaboration, in: Enterprise AI, Springer, 2025, pp. 211–243.

[9] T. Poell, D. Nieborg, J. van Dijck, Platformisation, Internet Policy Review 8 (2019) 1–13. URL: https://policyreview.info/node/1425. doi:10.14763/2019.4.1425.

[10] T. Flew, Regulating Platforms, Polity Press, Cambridge, UK, 2021.

[11] European Parliament and Council of the European Union, 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), EUR-Lex, 2022. URL: https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng, official Journal of the European Union, L 277, 27.10.2022, p. 1–102.

[12] M.-T. Sekwenz, R. Gsenger, The digital services act: Online risks, transparency and data access, in: Digital Decade: How the EU Shapes Digitalisation Research, volume 3, Nomos Verlagsgesellschaft mbH & Co. KG, 2025, pp. 115–140. URL: https://www.nomos-elibrary.de/de/10.5771/9783748943990-115/the-digital-services-act-online-risks-transparency-and-data-access. doi:10.5771/9783748943990-115, eISBN: 978-3-7489-4399-0.

[13] E. Commission, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj, legislative Body: CONSIL, EP.

[14] H. Hanif, J. Constantino, M.-T. Sekwenz, M. van Eeten, J. Ubacht, B. Wagner, Y. Zhauniarovich,

Navigating the EU AI Act Maze using a Decision-Tree Approach, ACM J. Responsib. Comput. (2024). URL: https://doi.org/10.1145/3677174. doi:10.1145/3677174, just Accepted.

[15] M.-T. Sekwenz, R. Gsenger, S. Dahlgren, B. Wagner, Doing Audits Right? The Role of Sampling and Legal Content Analysis in Systemic Risk Assessments and Independent Audits in the Digital Services Act, 2025. URL: http://arxiv.org/abs/2505.03601. doi:10.48550/arXiv.2505.03601, arXiv:2505.03601 [cs].

[16] J.-C. Plantin, C. Lagoze, P. N. Edwards, C. Sandvig, Infrastructure studies meet platform studies in the age of google and facebook, New Media & Society 20 (2018) 293–310. URL: https://doi.org/10.1177/1461444816661553. doi:10.1177/1461444816661553.

[17] Microsoft, Introducing bing generative search, Microsoft Bing Blogs, 2024. URL: https://blogs.bing.com/search/July-2024/generativesearch, accessed: 2026-01-16.

[18] R. Jahangir, Eu weighs regulating openai's chatgpt under the dsa. what does that mean?, TechPolicy.Press, 2025. URL: https://www.techpolicy.press/eu-weighs-regulating-openais-chatgpt-under-the-dsa-what-does-that-mean/, accessed: 2026-01-16.

[19] DelegatedRegulation2023, Delegated Regulation supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council, by laying down rules on the performance of audits for very large online platforms and very large online search engines, 2023. URL: file:///Users/mariesekwenz/Downloads/C_2023_6807_1_EN_ACT_part1_v2_6zoLEmP6RODW1Ujvl3uOmUHWPI_99436-3.pdf.

[20] C. Ling, I. AbuHilal, J. Blackburn, E. De Cristofaro, S. Zannettou, G. Stringhini, Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–24. URL: https://dl.acm.org/doi/10.1145/3449155. doi:10.1145/3449155.

[21] J. Kübler, M.-T. Sekwenz, F. Rachinger, A. König, R. Gsenger, E. Pírková, B. Wagner, M. C. Kettemann, M. Krennerich, C. Ferro, The 2021 german federal election on social media: Analysing electoral risks created by twitter and facebook, in: Proceedings of the 56th Annual Hawaii International Conference on System Sciences (HICSS 2023), IEEE, 2023, pp. 4036–4045. URL: https://hdl.handle.net/10125/103124. doi:10.24251/HICSS.2023.493.

[22] S. Biswas, A. Erlei, U. Gadiraju, Belief updating and delegation in multi-task human-ai interaction: Evidence from controlled simulations, arXiv preprint arXiv:2602.01986 (2026).

Acknowledgements

This research was part of RESOCIAL, funded by the Dutch Research Agenda, grant agreement NWA.1540.21.001. We gratefully acknowledge the contributions of the consortium and cooperation partners who supported the project through workshops, discussions, and feedback on interim results and outputs. Authors are partially supported by the TU Delft AI Initiative, the Model-driven Decisions Lab (MoDDL), the Convergence Flagship *ProtectMe*, and the Robust LTP *GENIUS* ICAI Lab. This publication was supported by the Weizenbaum Institute (grant number 16DII141), funded by the Federal Ministry of Research, Technology and Space (BMFTR) and the State of Berlin.