Human Oversight-by-Design for Accessible Generative IUIs

Blessing Jerry[1,*,†], Lourdes Moreno[1,†] and Paloma Martínez[1,†]

[1]Computer Science and Engineering Department, Universidad Carlos III de Madrid, Av. Universidad 30, 28911 Madrid, Spain

Abstract

LLM-generated interfaces are increasingly used in high-consequence workflows (e.g., healthcare communication), where how information is presented can impact downstream actions. These interfaces and their content support human interaction with AI-assisted decision-making and communication processes and should remain accessible and usable for people with disabilities. Accessible plain-language interfaces serve as an enabling infrastructure for meaningful human oversight. In these contexts, ethical and trustworthiness risks, including hallucinations, semantic distortion, bias, and accessibility barriers, can undermine reliability and limit users' ability to understand, monitor, and intervene in AI-supported processes. Yet, in practice, oversight is often treated as a downstream check, without clear rules for when human intervention is required or who is accountable. We propose oversight-by-design: embedding human judgment across the pipeline as an architectural commitment, implemented via escalation policies and explicit UI controls for risk signalling and intervention. Automated checks flag risk in generated UI communication that supports high-stakes workflows (e.g., readability, semantic fidelity, factual consistency, and standards-based accessibility constraints) and escalate to mandatory Human-in-the-Loop (HITL) review before release when thresholds are violated, or uncertainty is high. Human-on-the-Loop (HOTL) supervision monitors system-level signals over time (alerts, escalation rates, and compliance evidence) to tune policies and detect drift. Structured review feedback is translated into governance actions (rule and prompt updates, threshold calibration, and traceable audit logs), enabling scalable intervention and verifiable oversight for generative UI systems that support high-stakes workflows.

Keywords

user interfaces, human oversight, human-in-the-loop, AI governance, accessibility, LLM safety

1. Introduction

Large language models (LLMs) are rapidly being integrated into systems that operate in high-stakes domains such as healthcare, legal services, and public services. In these contexts, failures are not merely technical errors, but can threaten human safety, autonomy, and rights. For Intelligent User Interfaces (IUIs) that generate interface content and interaction cues, human oversight is not only a governance requirement but an interaction design problem: the system must expose risk, uncertainty, and intervention pathways through UI-level controls [1, 2, 3, 4]. Accessibility and plain language serve a dual role: they are quality and compliance criteria for generated interface content, and they are prerequisites for enabling diverse reviewers to effectively perform oversight and intervention tasks. Regulatory frameworks such as the EU AI Act require appropriate human oversight for high-risk AI systems to reduce risks to health, safety, and fundamental rights [5]. However, despite growing consensus on its importance, human oversight remains poorly specified in system design practice, and the Act provides limited guidance on how to implement effective oversight in concrete system architecture [6, 7, 8]. This gap is particularly critical for generative AI systems deployed in high-stakes domains, where the primary risks arise from the underlying decision and operational processes, and where interface design critically shapes whether humans can effectively understand, monitor, and intervene in those processes, whose failures may have immediate and severe consequences. In many deployed AI systems, oversight is implemented as a single review step or a nominal approval interface,

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus

*Corresponding author.

†These authors contributed equally.

bjerry@pa.uc3m.es (B. Jerry); lmoreno@inf.uc3m.es (L. Moreno); pmf@inf.uc3m.es (P. Martínez)

0009-0003-3628-3911 (B. Jerry); 0000-0002-9021-2546 (L. Moreno); 0000-0003-3013-3771 (P. Martínez)

offering limited visibility into system behaviour and limited intervention authority [9, 10, 11, 12]. Such designs can reduce intervention authority and visibility into system behaviour, and may foster over-reliance rather than effective control [11, 13, 14]. This gap between regulatory intent and technical implementation motivates our work.

We argue that effective human oversight requires architectural commitment across multiple system layers. Rather than treating HITL as a final validation gate, oversight mechanisms must be embedded within the system, automatically activated when risk signals exceed thresholds, and capable of supporting not only direct intervention (HITL) but also monitoring (HOTL) and governance through policy refinement [7, 15, 16, 17]. Effective oversight should ensure that decisions are logged with requirement identifiers, metric evidence, and rationales, and that human feedback is systematically used to refine system behaviour.

This paper presents a preliminary, methodological contribution: a multi-layer oversight architecture for generative accessible UIs that support high-stakes communication workflows. We instantiate the design in a healthcare communication setting, with accessible post-consultation medication instructions tailored to user profiles, including individuals with cognitive disabilities and low vision, integrating deterministic rules, LLM-based generation, automated evaluation metrics, and structured human feedback into an end-to-end auditable pipeline. The proposed oversight mechanism is domain-agnostic [17, 18]. We do not report user study results; instead, we provide an implemented architecture that operationalizes meaningful human oversight in practice. We provide:

- An end-to-end oversight architecture that operationalizes escalation-driven HITL and HOTL across generation, evaluation, and governance.
- A traceability scheme linking requirements, metric evidence, escalation decisions, and human rationales.
- A structured feedback loop that turns review outcomes into governance actions (rule and prompt updates, threshold calibration, and audit logs).

2. Related work

Human oversight has been studied across the human-automation interaction and HCI, commonly distinguishing direct intervention (Human-in-the-Loop, HITL) from supervisory control (Human-on-the-Loop, HOTL). Parasuraman et al. formalize levels of automation and show how misaligned allocation of control can harm performance and accountability [1]. Human factors research also highlights automation bias and loss of situation awareness as recurring failure modes when humans supervise complex systems [1, 19, 20, 21, 22].

In HCI, Amershi et al. provide widely used guidelines for Human–AI Interaction, emphasizing timing, transparency, and user control [23]. However, these guidelines primarily address interaction-level recommendations and leave open how to implement oversight as an enforceable mechanism across multi-stage generative pipelines [17, 24]. Related work on supervisory control shows that intervention quality degrades under high workload and limited transparency [13], and that explanation/control interfaces can increase over-reliance rather than calibrated trust [11, 14].

Recent syntheses of human–AI collaboration further highlight that many deployed systems still rely on limited interaction patterns and weak forms of collaborative control. Gomez et al. identify recurring structures of human–AI interaction in decision-support pipelines and show that collaboration often remains asymmetric and poorly integrated into workflow design [25]. This reinforces the need for architectural approaches that explicitly structure intervention, supervision, and continuous improvements across system pipelines, rather than treating human involvement as a single interaction layer.

Accessibility standards and guidance (WCAG 2.2 [26], EN 301 549 [27], ISO 24495-1 [28], W3C COGA [29]) stress that information must be understandable, not only correct, including information presented to human reviewers and supervisors during oversight and escalation activities, aligning with Cognitive Load Theory's account of how complexity and poor structure impair comprehension in safety-critical

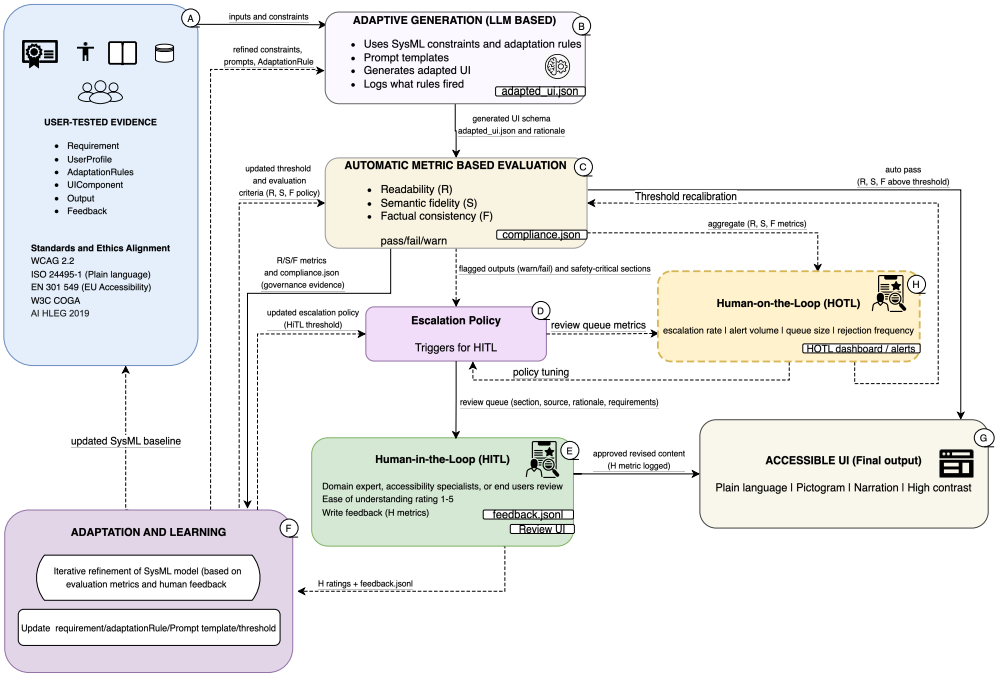

Figure 1: End-to-end architecture for accessible UI generation with layered human oversight

contexts [30, 31, 32]. Building on these foundations, our work operationalizes escalation-driven HITL and HOTL as an auditable architecture that links risk signals, intervention points, and structured feedback to ongoing governance [16, 17]. Automated metrics are frequently used to assess generated outputs. Semantic similarity metrics (e.g., BERTScore) [33], factual consistency measures (e.g., FactCC) [34], and readability indices can provide useful signals, but they are often insufficient proxies for comprehension, particularly for cognitively vulnerable users [30, 35, 36, 37]. We therefore treat metrics as risk detectors that trigger escalation rather than as final arbiters.

Concretely, this work extends our prior model-driven architecture for LLM-driven accessible interfaces [38], featuring SysML v2 traceability and standards-aligned templates, with escalation-driven HITL and HOTL oversight, and governance mechanisms.

3. System architecture overview

We introduce a design-driven architecture for generating accessible user interfaces and UI content using large language models (LLMs). The approach is model-based and aligns with standards: it employs structured user profiles, declarative adaptation rules, and validated prompt templates to create multimodal accessible outputs (e.g., plain-language text, pictograms, high-contrast layouts) with clear traceability to normative requirements (e.g., WCAG, EN 301 549, ISO 24495-1, W3C COGA).

Figure 1 summarizes the end-to-end pipeline. (A) User-tested evidence captures user profiles, requirements, adaptation rules, and UI components, and anchors them to standards. This layer acts as the source of truth for what must be satisfied and what evidence must be logged. In prior work, SysML v2 is used to maintain explicit traceability between user needs, adaptation rules, and normative requirements, enabling auditable transformations in generative accessible interfaces.

(B) Adaptive generation (LLM-based) produces a structured UI schema (e.g., headings, sections, modality choices, theming) guided by constraints derived from the requirement model and rule set. The goal is not free-form UI generation, but controlled generation bounded by (i) baseline accessible templates and (ii) profile-driven refinements (e.g., plain language, stepwise structuring, pictograms).

(C) Automatic metric-based evaluation computes checkpoint signals for generated UI communication, including readability, semantic fidelity, and factual consistency. These signals are logged as machine-

readable evidence (e.g., in a compliance.json artifact) and are used as risk indicators rather than final judgments. This is consistent with our POC pipeline that formalizes checkpoints R, S, F, H (R using Flesch-Szigriszt and Fernández–Huerta (Spanish) [39, 40], S using BERTScore [33], F using AlignScore [41], and H captured via Ease-of-Understanding (1–5) ratings with comments while logging evidence for auditability.

(D) Escalation policy translates risk signals into enforceable control: outputs that violate thresholds, contain safety-critical segments, or show high uncertainty are routed to mandatory human review prior to release. This is the architectural "commitment" that turns evaluation into governance, rather than leaving oversight as an afterthought.

(E) Human-in-the-Loop review is performed by the appropriate stakeholders (domain experts, accessibility specialists, and end users) using structured review instruments (e.g., ease-of-understanding ratings and component-level annotations). Feedback is recorded in a structured log (e.g., feedback.jsonl) linked to requirement IDs and UI nodes, enabling traceable corrective actions. The POC similarly captures human feedback as structured events and links it back to SysML requirement identifiers, closing the loop between human validation and model updates.

Finally, (F) Adaptation and learning aggregates metric evidence and human feedback to update governance artifacts (rules, prompt templates, thresholds, and requirement refinements), while preserving versioned traceability. We instantiate the architecture in healthcare communication (patient-facing interface content derived from clinical text), consistent with prior model-driven accessible UI generation work; however, the oversight mechanism (links risk signals, escalation, structured review, and governance updates) is domain-agnostic and can transfer to other high-stakes UI settings (e.g., legal or public services).

4. Oversight mechanism

Our oversight model separates intervention, supervision, and governance so responsibilities remain explicit and auditable.

4.1. Oversight Modalities

- **HITL (intervention).** HITL is activated by the escalation policy when automated signals are insufficient to guarantee safety or accessibility. Reviewers validate: (i) meaning preservation relative to the source, (ii) factual consistency for safety-relevant communication claims, and (iii) understandability for the target profile. Outputs include approval or rejection, targeted revision requests, and structured rationales linked to requirement and UI component identifiers. In the healthcare communication use case, HITL reviewers act as clinical communication reviewers and accessibility reviewers for post-consultation medication instructions. Escalation is triggered, for example, when (a) readability or plain-language scores fall below thresholds for a cognitive-impairment profile, (b) contradictions are detected between generated instructions and structured medication data (e.g., dosage, timing, or route of administration), or (c) warnings related to known contraindications or safety advice are missing. Reviewers may edit or request revisions to the generated instruction text, confirm or correct safety-relevant statements, and validate that the presentation and wording satisfy accessibility and understandability requirements for the intended user profile.
- **HOTL (supervision).** HOTL monitors the system's behavior over time rather than reviewing every output. Supervisors inspect aggregate signals such as alert volume, escalation rates, failure modes by requirement category, and compliance evidence trends, and use these to recalibrate thresholds and detect drift. In the healthcare communication setting, HOTL supervision focuses on escalation rates and failure patterns across user profiles (e.g., cognitive impairment or low-vision profiles) and across instruction types (e.g., single-drug versus multi-drug regimens). Drift is indicated, for example, by rising numbers of semantic mismatches between generated

instructions and structured medication data, or by repeated reviewer overrides for the same classes of accessibility or safety requirements.

- **Governance (continuous improvement).** Governance converts review outcomes into controlled updates to the system's "policy surface": adaptation rules, prompt templates, threshold settings, and requirement refinements. This mirrors the POC's trace-first approach in which evaluation evidence and feedback are tied to requirement IDs to support auditable updates. For the healthcare communication use case, governance actions include updating plain-language and accessibility adaptation rules, refining prompt templates for medication instruction generation, and recalibrating escalation thresholds based on recurring reviewer feedback for specific risk patterns (e.g., persistent readability violations for certain cognitive profiles or recurring safety-related omissions).

This healthcare communication use case is employed to ground the oversight architecture in a realistic, safety-sensitive communication scenario and does not imply that the system performs or replaces clinical decision-making.

4.2. Escalation policy

The escalation policy is explainable and auditable: it records which signals triggered review, which UI sections were flagged (including safety-relevant communication segments), and which stakeholder role is required for resolution. Human feedback is captured as structured, versioned events referencing the same identifiers used in the requirement model, enabling traceable governance actions and compliance reporting.

5. Trade-offs and Roles

Oversight must balance safety, scalability, and workload. The architecture avoids two extremes: (i) fully automated approval based on metrics alone, and (ii) universal manual review. Instead, it makes HITL selective and risk-proportional via escalation, while HOTL ensures continuous monitoring and calibration. Stakeholder roles map to evidence types:

- End users: provide comprehension-oriented feedback (ease-of-understanding, usability notes) that standards alone cannot guarantee.
- Domain experts: validate correctness and safety in escalated cases.
- Accessibility specialists: verify standards-based constraints and profile fit.
- System owners and governance: tune policy thresholds, review audit logs, and maintain traceability for compliance.

6. Discussion and work in progress

This work argues that meaningful oversight in generative IUIs is an architectural property: risk must be surfaced through UI-level signals, escalation must be enforceable, and decisions must be traceable to requirements and evidence. Our current contribution is methodological, demonstrating an implemented pipeline that integrates model-driven accessibility constraints with risk-based escalation and structured review.

Limitations include (i) the need for empirical validation of how automated signals align with human judgments across profiles, and (ii) the need for longitudinal monitoring to study drift and policy recalibration in deployment. Next steps include user studies with target populations and domain professionals, and evaluating whether checkpoint thresholds and escalation strategies generalize across user-interface settings that support high-stakes workflows.

This workshop paper focuses on oversight-by-design and does not replicate our prior model-driven accessible-interface pipeline, which provides SysML v2 traceability and standards-aligned templates

[38]. We also draw on an ongoing proof-of-concept in healthcare communication (unpublished) as an instantiation context.

Acknowledgments

This work has also been supported by grants PID2023-148577OB-C21 (Human-Centered AI: User Driven Adapted Language Models-HUMAN_AI) funded by MICIU/AEI/10.13039/501100011033 and by FEDER/UE.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to: Grammar and spelling check, paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Parasuraman, T. B. Sheridan, C. D. Wickens, A model for types and levels of human interaction with automation, IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans 30 (2000) 286–297. doi:10.1109/3468.844354.

[2] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human-Computer Interaction 36 (2020) 495–504. doi:10.1080/10447318.2020.1741118.

[3] Y. Zhang, Q. V. Liao, R. K. E. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*), 2020, pp. 295–305. doi:10.1145/3351095.3372852.

[4] S. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, J. W. Vaughan, I am not sure, but...: Examining the impact of large language models' uncertainty expression on user reliance and trust, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2024, pp. 822–835. doi:10.1145/3630106.3658941.

[5] European Parliament and Council of the European Union, 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), Regulation (EU) 2024/1689, Official Journal of the European Union, 2024. URL: http://data.europa.eu/eli/reg/2024/1689/oj, 12 July 2024.

[6] L. Enqvist, Human oversight in the EU artificial intelligence act: what, when and by whom?, Law, Innovation and Technology 15 (2023) 508–535. doi:10.1080/17579961.2023.2245683.

[7] M. Ho-Dac, B. Martinez, Human oversight of artificial intelligence and technical standardisation, arXiv preprint, 2024. arXiv:2407.17481.

[8] A. Hummel, H. Burden, S. Stenberg, J.-P. Steghöfer, N. Kühl, The EU AI act, stakeholder needs, and explainable AI: Aligning regulatory compliance in a clinical decision support system, arXiv preprint, 2025. arXiv:2505.20311.

[9] L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making?, International Journal of Human-Computer Studies 51 (1999) 991–1006. doi:10.1006/IJHC.1999.0252.

[10] K. L. Mosier, L. J. Skitka, S. Heers, M. Burdick, Automation bias: decision making and performance in high-tech cockpits, The International Journal of Aviation Psychology 8 (1997) 47–63. doi:10.1207/S15327108IJAP0801_3.

[11] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making, Proceedings of the ACM on Human-Computer Interaction 5 (2021) 1–21. doi:10.1145/3449287.

[12] B. Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022). URL: https://linkinghub.elsevier.com/retrieve/pii/S0267364922000292. doi:10.1016/j.clsr.2022.105681.

[13] M. M. Cummings, Man versus machine or man + machine?, IEEE Intelligent Systems 29 (2014) 62–69. doi:10.1109/MIS.2014.87.

[14] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, Does the whole exceed its parts? the effect of AI explanations on complementary team performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21), Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–16. doi:10.1145/3411764.3445717.

[15] S. Sterz, K. Baum, S. Biewer, H. Hermanns, A. Lauber-Rönsberg, P. Meinel, M. Langer, On the quest for effectiveness in human oversight: Interdisciplinary perspectives, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), Association for Computing Machinery, New York, NY, USA, 2024, pp. 2495–2507. doi:10.1145/3630106.3659051.

[16] D. Manheim, Oversight of unsafe systems via dynamic safety envelopes, arXiv preprint, 2018. arXiv:1811.09246.

[17] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20), Association for Computing Machinery, New York, NY, USA, 2020, pp. 33–44. doi:10.1145/3351095.3372873.

[18] Y. Artsi, V. Sorin, B. S. Glicksberg, P. Korfiatis, G. N. Nadkarni, E. Klang, Large language models in real-world clinical workflows: A systematic review of applications and implementation, Frontiers in Digital Health 7 (2025). doi:10.3389/fdgth.2025.1659134.

[19] M. R. Endsley, Toward a theory of situation awareness in dynamic systems, Human Factors 37 (1995) 32–64. doi:10.1518/001872095779049543.

[20] M. Schemmer, N. Kühl, C. Benz, G. Satzger, On the influence of explainable AI on automation bias, arXiv preprint, 2022. arXiv:2204.08859.

[21] J. M. Echterhoff, Y. Liu, A. Alessa, J. McAuley, Z. He, Cognitive bias in decision-making with LLMs, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 12640–12653. URL: https://aclanthology.org/2024.findings-emnlp.739/. doi:10.18653/v1/2024.findings-emnlp.739.

[22] H. Li, S. Aral, Human trust in AI search: A large-scale experiment, arXiv preprint, 2025. arXiv:2504.06435.

[23] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-AI interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19), Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–13. doi:10.1145/3290605.3300233.

[24] F. Ferrari, J. van Dijck, A. van den Bosch, Observe, inspect, modify: Three conditions for generative AI governance, New Media & Society 27 (2025) 2788–2806. doi:10.1177/14614448231214811.

[25] C. Gomez, S. M. Cho, S. Ke, C.-M. Huang, M. Unberath, Human-ai collaboration is not very collaborative yet: A taxonomy of interaction patterns in ai-assisted decision making from a systematic review, Frontiers in Computer Science 6 (2024) 1521066. URL: https://www.frontiersin.org/articles/10.3389/fcomp.2024.1521066. doi:10.3389/fcomp.2024.1521066, section: Human-Media Interaction. Published online 6 January 2025.

[26] W3C Web Accessibility Initiative, Web Content Accessibility Guidelines (WCAG) 2.2, W3C Recommendation, World Wide Web Consortium, 2023. URL: https://www.w3.org/TR/WCAG22/.

[27] ETSI, CEN, CENELEC, EN 301 549 V3.2.1: Accessibility requirements for ICT products and services, European Standard EN 301 549 V3.2.1, European Telecommunications Standards Institute, 2021. URL: https://www.etsi.org/deliver/etsi_en/301500_301599/301549/03.02.01_60/en_301549v030201p.pdf.

[28] International Organization for Standardization, ISO 24495-1:2023 Plain language — Part 1: Gov-

erning principles and guidelines, International Standard ISO 24495-1:2023, ISO, 2023. URL: https://www.iso.org/standard/78907.html.
[29] W3C Web Accessibility Initiative, Making Content Usable for People with Cognitive and Learning Disabilities (COGA), W3C Working Draft, World Wide Web Consortium, 2021. URL: https://www.w3.org/TR/coga-usable/.
[30] J. Sweller, Cognitive load during problem solving: Effects on learning, Cognitive Science 12 (1988) 257–285. doi:10.1016/0364-0213(88)90023-7.
[31] D. Lyell, F. Magrabi, E. Coiera, The effect of cognitive load and task complexity on automation bias in electronic prescribing, Human Factors: The Journal of Human Factors and Ergonomics Society 60 (2018) 1008–1021. doi:10.1177/0018720818781224.
[32] L. Moreno, H. Petrie, P. Martínez, R. Alarcón, Designing user interfaces for content simplification aimed at people with cognitive impairments, Universal Access in the Information Society 23 (2024) 99–117. doi:10.1007/s10209-023-00986-z.
[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.
[34] W. Kryściński, B. McCann, C. Xiong, R. Socher, Evaluating the factual consistency of abstractive text summarization, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9332–9346. URL: https://aclanthology.org/2020.emnlp-main.750/. doi:10.18653/v1/2020.emnlp-main.750.
[35] P. Martínez, A. Ramos, L. Moreno, Exploring large language models to generate easy to read content, Frontiers in Computer Science 6 (2024) 1394705. doi:10.3389/fcomp.2024.1394705.
[36] R. Alarcón, L. Moreno, P. Martínez, J. A. Macías, EASIER system: Evaluating a spanish lexical simplification proposal with people with cognitive impairments, International Journal of Human–Computer Interaction 40 (2024) 1195–1209. doi:10.1080/10447318.2022.2134074.
[37] A. Säuberli, F. Holzknecht, P. Haller, S. Deilen, L. Schiffl, S. Hansen-Schirra, S. Ebling, Digital comprehensibility assessment of simplified texts among persons with intellectual disabilities, in: CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–11. doi:10.1145/3613904.3642570.
[38] B. Jerry, L. Moreno, V. Francisco, R. Hervas, LLM-driven accessible interface: A model-based approach, 2025. Accepted for publication; to appear in the EMUXAI 2025 Workshop on Engineering Methods for HCI and UX in AI-Driven Systems (INTERACT 2025 Workshop), Lecture Notes in Computer Science (LNCS), Springer.
[39] F. S. Pazos, Sistemas predictivos de legibilidad del mensaje escrito: fórmula de perspicuidad, Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain, 1992. URL: https://hdl.handle.net/20.500.14352/62699, doctoral dissertation, Facultad de Ciencias de la Información.
[40] J. F. Huerta, Medidas sencillas de lecturabilidad, Consigna (1959) 29–32.
[41] Y. Zha, Y. Yang, R. Li, Z. Hu, Alignscore: Evaluating factual consistency with a unified alignment function, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics, 2023, pp. 11328–11348. URL: https://aclanthology.org/2023.acl-long.634/. doi:10.18653/v1/2023.acl-long.634.