Calibrating Clinician Trust in Explainable AI Summarization for Obstetrics: A Sociotechnical Design Study of the MedGPT Portugal Use Case

Filipe Bernardi[1,*], Daniel Rodrigues[2], Rute Almeida[1], Vinicius Lima[1], Isabela Guerche[1], Julio Souza[3], Ricardo Correia[2], Luis Conceicao[3], Goreti Marreiros[3] and Alberto Freitas[1]

[1]*RISE-Health, MEDCIDS, Faculty of Medicine, University of Porto, Portugal*
[2]*VirtualCare, LDA, Porto, Portugal*
[3]*GECAD/LASI, ISEP – Polytechnic of Porto, Portugal*

Abstract

Large language models can reduce documentation burden by generating clinical summaries from electronic health records, but obstetric use cases are safety-critical because fluent text may hide omissions, temporal inconsistencies, or unsupported claims. We present a condensed sociotechnical design study for the MedGPT-ObsCare Portugal use case, treating trust as a calibration target rather than a variable to maximize. The paper contributes: (i) a multidimensional design space for explainable obstetric summarization; (ii) four explanation modes (baseline, narrative rationale, evidence-anchored, and layered progressive disclosure); (iii) an executable pre-deployment evaluation protocol focused on reliance and verification behavior; and (iv) governance requirements for regulated settings, including role-based access control, auditability, and versioned knowledge assets. The contribution is methodological and pre-deployment oriented, providing a reusable framework for staged pilot evaluation.

Keywords

Explainable AI, Clinical summarization, Trust calibration, Human-in-the-loop, Obstetrics

1. Introduction

Electronic health records (EHRs) increase data availability but also create a burden for navigation and synthesis [1]. In obstetrics, clinicians must quickly combine longitudinal information from structured entries, free-text notes, laboratory data, and imaging findings, often under time pressure. Artificial intelligence (AI)-based summarization can support situational awareness, yet in safety-critical domains errors are costly: plausible language may conceal missing evidence, fabricated details, or outdated clinical context [2, 3, 4].

For this reason, we frame explainable clinical summarization as a trust calibration problem and position our contribution as methodological and pre-deployment rather than empirical validation. Beyond describing the MedGPT-ObsCare Portugal use case, the paper organizes a reusable framework that links interface design, reliance behavior, and governance in safety-critical summarization workflows. Specifically, it contributes a multidimensional design space, four explanation modes for obstetric summarization interfaces, and an executable pre-deployment evaluation protocol focused on reliance and verification behavior under realistic review tasks. It also specifies governance requirements for regulated settings, including role-based access control, auditability, and versioned knowledge assets. This matters beyond Portugal because the framework is intended to support context-sensitive pilot planning and accountable adoption of clinical summarization systems across healthcare environments.

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil
*Corresponding author.
fbernardi@med.up.pt (F. Bernardi)
0000-0002-9597-5470 (F. Bernardi); 0000-0002-0777-1540 (D. Rodrigues); 0000-0001-7755-5002 (R. Almeida); 0000-0002-2467-358X (V. Lima); 0009-0006-9468-0247 (I. Guerche); 0000-0002-8576-1903 (J. Souza); 0000-0002-3764-5158 (R. Correia); 0000-0003-3454-4615 (L. Conceicao); 0000-0003-4417-8401 (G. Marreiros); 0000-0003-2113-9653 (A. Freitas)

2. Background and Related Work

2.1. Trust, factuality, and automation bias

Clinical summarization with large language models has progressed rapidly, but three concerns remain central in healthcare settings: factual faithfulness, omission of critical details, and over-reliance on fluent but weakly verifiable text [2, 5, 6, 4]. In obstetrics, these risks are amplified by longitudinal records, context-sensitive interpretation, and frequent time pressure. Prior studies also show that interface framing can alter clinician reliance independently of model quality, reinforcing the need to treat trust as a calibration problem rather than a confidence-maximization problem [7, 8, 9, 10].

In this framing, trust is *appropriate reliance* under uncertainty, where both over-trust and under-trust are safety-relevant failure modes. For clinical summarization, the practical unit of failure is a textual clinical claim whose provenance may be costly to verify during routine work [2, 7]. Human-in-the-loop safeguards therefore depend not only on model output quality, but also on whether the interface supports proportionate verification in low- and high-risk situations [8, 9].

2.2. Why explainability alone is insufficient

Explainability is often proposed as the main response to opaque model behavior, but explanation quality is only one component of a broader sociotechnical system [11, 9]. Rationale-only explanations may improve readability while increasing persuasive trust, whereas evidence-anchored designs provide stronger epistemic support by linking claims to inspectable sources and uncertainty cues, typically at higher interaction cost. In safety-critical summarization, these trade-offs must be evaluated in relation to workflow demands, role responsibilities, and governance controls.

Current literature has advanced explainability in healthcare AI, trust calibration, human-in-the-loop decision support, and LLM-based clinical summarization, but these strands are still mostly treated in parallel rather than as an integrated pre-deployment design problem [11, 9, 6, 4, 10]. The gap is a workflow-aware framework that explicitly connects explanation design, reliance and verification behavior, and governance constraints before deployment. Accordingly, this paper contributes a calibration-oriented, pre-deployment methodological framework—not a new model—to structure pilot evaluation and accountable implementation. The next section details the methods used to operationalize this framework in the MedGPT-ObsCare context.

3. Context and Methods

3.1. Study design and scope

This paper follows a theory-informed, design-oriented method and is intentionally limited to the pre-deployment stage [13, 12]. We do not claim clinical impact or model superiority. Instead, we specify a testable design framework in which key parameters are operationalized as observable behaviors in realistic obstetric review tasks.

The analysis integrates three evidence sources: (i) published findings on summarization faithfulness, trust calibration, and cognitive load [5, 11, 9, 6]; (ii) MedGPT-ObsCare requirements and pilot constraints[1] [14]; and (iii) structured scenario-based reasoning aligned with clinical workflow in Portuguese obstetric services. This approach is suitable for safety-critical contexts where premature claims of effectiveness would be methodologically weak.

Clinician involvement at this stage was limited to informing design discussions and contextual validation of the ObsCare use case. Broader and more representative clinician participation is planned for staged pilot phases, including participatory refinement of the risk-adaptive explanation policy.

[1] https://virtualcare.pt/medgpt-medical-gpt-revolutionizing-healthcare-with-ethical-ai/

3.2. MedGPT-ObsCare use case

MedGPT-ObsCare is treated as a design artefact and pilot pathway. The intended functionality is to generate concise narrative summaries that combine structured and unstructured EHR data across pregnancy trajectories, supporting handovers, consultations, and rapid contextualization [3, 4]. The system is advisory only, and all outputs require professional validation before clinical reuse [10]. Within this paper, the MedGPT-ObsCare Portugal case functions as a design context used to ground requirements, interface choices, and governance assumptions. The primary contribution, however, is a generalizable methodological trust-calibration framework for pre-deployment evaluation that can be transferred beyond this specific case.

From an architectural perspective, the use case assumes retrieval-augmented generation and provenance linkage, enabling claims to be traced back to supporting records. This is consistent with current recommendations for reducing hallucination exposure in high-risk text generation settings [5, 6, 15]. However, retrieved evidence alone does not guarantee calibrated use if links are hidden, overwhelming, or poorly integrated with clinician verification routines.

3.3. Trust calibration as an operational objective

We define trust calibration as alignment between subjective trust and observed behavior. In this study, behavior is represented by reliance choices (accept, edit, discard), verification actions (opening evidence and cross-checking key claims), corrective actions (edits, overrides, escalations), and workload-sensitive interaction patterns (time and attention allocation). This operationalization allows evaluation of explainability as a behavioral risk control rather than a purely perceptual feature. Table 1 maps these dimensions to behaviors.

Table 1
Mapping of design dimensions to expected clinician behaviors

Design dimension	Behavior at risk	Behavior to promote
Low epistemic strength	Acceptance without checking	Targeted verification before acceptance
High persuasion risk	Reliance on writing fluency	Reliance on evidence quality
High verification cost	Skipped checks in urgent tasks	Selective drill-down and prioritization
Low user control	Passive consumption	Active inspection, editing, and override
Low traceability	Weak accountability	Auditable decision pathway

4. Design Space and Explanation Modes

4.1. Design dimensions

We compare explanation strategies across seven dimensions: epistemic strength, verification cost, granularity, traceability, persuasion risk, user control, and cognitive load. The dimensions are deliberately tension-based: improvements in one axis may degrade another [9]. This representation helps explain why one universal interface is unlikely to be safe across all obstetric contexts. Table 2 summarizes these trade-offs.

Table 2: Expected trade-offs across explanation modes

Mode	Epistemic support	Verification cost	Persuasion risk	Workflow fit
E0	Low	Low	Medium	High speed, weak auditability
E1	Low–Medium	Low	High	Fast, may induce over-reliance
E2	High	High	Low	Strong verification and traceability
E3	High (on demand)	Medium	Low–Medium	Balances speed and evidence access

4.2. Modes E0–E3

We operationalize four modes: E0 (Baseline), summary text only; E1 (Narrative rationale), short natural-language justification without direct links; E2 (Evidence-anchored), claim-level links to supporting excerpts; and E3 (Layered progressive disclosure), concise summary first with on-demand rationale and evidence drill-down [11].

In low-risk routine encounters, E0/E1 can support speed but may require stronger organizational safeguards. In higher-risk or uncertain cases, E2/E3 offer stronger epistemic support by making evidence and rationale inspectable. E3 is particularly relevant when teams need to balance rapid reading with selective deep verification.

5. Pre-Deployment Evaluation Protocol

5.1. Research question and hypotheses

Research question: *How do explanation modes affect trust calibration, verification behavior, and cognitive workload in obstetric summarization tasks?*

Hypotheses for staged pilot testing:

- **H1:** E2 and E3 reduce inappropriate reliance compared with E0 and E1.
- **H2:** E2 increases verification time and perceived workload.
- **H3:** E3 preserves calibration gains while reducing E2 workload.
- **H4:** E1 can increase persuasive trust without equivalent reliability gains.

5.2. Tasks and scenario design

The protocol uses clinically realistic scenarios spanning pre-conception, pregnancy, and puerperium. Encounter types include routine outpatient visits, inpatient follow-up, and emergency triage. In each scenario, clinicians interact with a single explanation mode and complete a structured review task.

To stress-test calibration, the scenario set should include seeded failures that reflect known model risks: omitted high-priority findings, unsupported claims presented as facts [2, 3], temporal inconsistencies, and ambiguity around uncertainty or confidence cues.

This design enables direct observation of when users detect, correct, or miss errors under each explanation strategy.

5.3. Measures, instrumentation, and analysis

We combine expert assessment with behavioral and human-factor metrics: summary quality (factual consistency, key-concept coverage, omission severity), reliance outcomes (accept/edit/discard and override rates), verification traces (evidence opens, dwell time, source switching), workload indicators

(including the NASA Task Load Index, NASA-TLX [16]), and trust-reliance alignment (subjective trust versus observed action).

Human-in-the-loop instrumentation logs interaction events together with model/prompt/retrieval versions [12, 13]. This allows traceable linkage between explanation design decisions and downstream behavior. Mixed-effects analysis can estimate the influence of explanation mode, scenario type, and clinician profile on calibration outcomes.

This protocol is executable because seeded failures are paired with explicit reliance choices, verification traces, and workload indicators in each reviewed scenario. All interaction events are captured in logs with versioning of model, prompt, and retrieval assets, enabling reproducible linkage between observed behavior and specific system states. These elements support consistent case-level assessment now, while effectiveness claims remain reserved for future pilot phases.

5.4. Interpretation logic for calibration outcomes

To keep the protocol decision-oriented, outcome interpretation should separate *summary correctness* from *clinician reliance behavior*. This yields four practically relevant patterns: calibrated acceptance (correct summary accepted), calibrated rejection (flawed summary rejected or corrected), over-reliance (flawed summary accepted with insufficient checks), and under-reliance (correct summary rejected without substantive justification). Reporting these patterns side by side is more informative than aggregate trust scores alone, because it shows whether explanation modes improve true safety-relevant behavior.

For pilot governance, we recommend pre-specifying progression criteria between phases (sandbox, shadow, bounded use) using this behavioral logic. Examples include upper bounds for acceptance of seeded high-severity errors, minimum verification activity in high-risk scenarios, and explicit documentation quality for overrides and escalations. These criteria should be treated as adaptive guardrails, not static pass/fail metrics: thresholds can be revised by the clinical governance board as incident patterns and workflow constraints become clearer during iterative testing.

This interpretation layer also improves comparability across updates. When model checkpoints, retrieval assets, or prompt templates change, teams can evaluate whether apparent quality gains are accompanied by safer reliance behavior, or merely by increased persuasive acceptance. In this way, calibration analysis functions as a bridge between interface design, model lifecycle management, and real-world accountability.

6. Safety and Governance Requirements

In regulated settings, the quality of explanations is insufficient without governance controls [10]. The implementation pathway should include: role-based access control with least-privilege defaults, immutable audit trails for generated outputs and clinician actions, version control for prompts/models/retrieval assets, escalation logic for high-risk contexts, and explicit assignment of accountability to licensed professionals [12]. These mechanisms convert trust calibration into an auditable operational process. They also support post-incident investigation and safe, iterative updates, which are essential when generative behavior changes following revisions to the model or knowledge base.

7. Pilot deployment strategy and risk-adaptive explanation policy

To reduce deployment risk, we recommend a phased pilot pathway that gradually increases ecological complexity while preserving traceability and governance oversight. The pathway comprises three stages: sandbox evaluation with synthetic or de-identified cases to assess usability, evidence navigation, and error-detection behavior; shadow mode, in which the system runs in parallel with routine documentation without affecting official records, enabling comparison with standard practice and identification of

latent hazards; and bounded pilot use, in which summaries may support selected workflows under explicit supervision and escalation constraints.

Across these stages, a practical deployment policy is to adapt the explanation mode to the case's risk and uncertainty, rather than applying a single fixed interface to all encounters. The objective is not strict automation of interface selection, but transparent alignment between epistemic needs and interaction burden. Table 3 summarizes this policy. For reproducible pilot analysis, each evaluated case should preserve scenario metadata, model and retrieval versions, explanation mode, clinician interaction traces, and final adjudicated outcome. Reporting should include both aggregate performance and failure-specific analyses, especially high-severity omissions, inappropriate acceptance of flawed outputs, and overrides of correct outputs. Pilot readiness also depends on reconstructing how each summary was produced, reviewed, and finalized using versioned telemetry and audit logs, thereby supporting periodic reviews of configuration changes and safety-relevant reliance patterns across pilot phases.

Table 3: Suggested explanation policy by clinical context

Context	Domain Risk	Preferred mode	Mandatory control
Routine follow-up with stable trajectory	Time pressure and minor omissions	E1 or E3	Quick checklist confirmation
Inpatient review with multi-day evolution	Temporal inconsistency across notes	E3	Drill down on key claims
Emergency assessment with incomplete data	Over-reliance under uncertainty	E2 or E3	Validate statements
Complex comorbidity or conflicting data	Ambiguity and contradictory evidence	E2	Senior review with documented override

8. Discussion and Limitations

The main contribution is methodological: a practical bridge between explainable-AI interface choices and measurable clinician behavior in obstetric summarization [9, 11]. By centering calibration rather than persuasion, the framework helps avoid two common failure patterns: opaque, speed-first summaries that encourage silent over-reliance, and heavy-evidence views that are theoretically robust but operationally ignored [7, 8].

Limitations remain significant. This manuscript does not provide effectiveness outcomes, and behavior in pilot scenarios may diverge from in-situ use. Evidence-rich interfaces can also shift workload from the model to the clinician, creating new friction points during high-pressure shifts [1]. Finally, transferability to other specialties depends on local documentation cultures and the maturity of governance. Limitations also include limited scenario realism, learning effects from repeated exposure, a mismatch between seeded and real-world failures, and limited generalizability across workflows [9].

A further practical limitation is that calibration quality depends not only on explanation design but also on the surrounding workflow. If verification opportunities are poorly timed relative to ward interruptions, clinicians may skip checks even when evidence links are available [7]. Conversely, if interfaces force excessive checking in low-risk encounters, teams may develop workarounds that reduce adherence to the intended protocol. For this reason, pilot interpretation should explicitly separate model-quality effects from workflow-friction effects and should report both average behavior and context-conditioned behavior.

From an implementation perspective, the most robust strategy is to treat explanation modes as configurable risk controls rather than fixed product features. Local governance boards can define default modes by context, set mandatory checks for selected clinical statements, and refine thresholds using observed incident patterns [10]. This governance loop creates a path to incremental safety gains without claiming full automation readiness, and helps maintain clinician agency as the central safeguard in AI-assisted summarization [12].

An additional limitation is that the present framework is primarily clinician-facing and does not yet operationalize patient-facing governance in depth. Future pilot governance and evaluation should therefore include transparent patient communication about AI-assisted documentation, clear opt-out procedures, and AI literacy support for patients and families, consistent with evidence on pregnant patients' views of AI in clinical care [17] and obstetricians' and midwives' perspectives on epistemic reliance in AI-supported decision-making [18]. Consistent with documented MedGPT patient-side requirements, these elements can be incorporated through multilingual access, educational content, and structured patient feedback mechanisms.

Next steps are to run staged pilot studies with representative clinicians, prioritize seeded-error stress tests, and iteratively refine escalation criteria, the depth of evidence presentation, and review prompts.

9. Conclusion

This late-breaking version extends the original full-paper argument while preserving pre-deployment scope. Explainable obstetric summarization is framed as a sociotechnical calibration problem rather than as a standalone model feature. The expanded design space, behavior-centered protocol, and governance controls provide a realistic pathway for safe pilot evaluation of MedGPT-ObsCare before broad deployment. Future pilot governance should also extend this pathway with patient-facing transparency, opt-out, and AI literacy components aligned with existing MedGPT multilingual information and feedback requirements.

Acknowledgements

This work was developed within the framework of the MedGPT project (ITEA-2023-23020), funded by the European Regional Development Fund (ERDF) and Portugal 2030 through the Innovation and Digital Transition Program, under project number NORTE2030-FEDER-01414600. The authors also acknowledge institutional support through tenure-funded positions at the Faculty of Medicine of the University of Porto for R.A. (2023.13694.TENURE.056).

Declaration on Generative AI

During the preparation of this work, the author(s) used Generative AI tools for language editing and formatting support only. All technical content, claims, and final revisions were reviewed and validated by the author(s), who take full responsibility for the publication's content.

References

[1] Y. Pinevich, K. Clark, K. H. Klement, V. Herasevich, Interaction time with electronic health records: a systematic review, Appl. Clin. Inform. 12(4) (2021) 788–799. doi:10.1055/s0041-1733909.

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55(12) (2023) Article 248, 1–38. doi:10.1145/3571730.

[3] T. C. Draper, K. Lamb-Riddell, C. Cox, N. Lewis, J. K. Archer, A. V. Rowe, L. Kravariti, S. Edwards, A. B. Thiemann, D. R. Fulton, Clinical AI scribes in primary care: accuracy, error severity and implications for clinical practice, BMJ Digital Health & AI 1(1) (2025) e000092. doi:10.1136/bmjdhai-2025-000092.

[4] C. Y. K. Williams, A. Bains, S. M. Wangensteen, S. J. Kazemi, N. S. Maniar, M. M. Serina, M. Vyas, A. K. A. Kung, R. J. Hussain, A. M. Chen, S. Loesberg, H. A. Tseng, N. V. Kulkarni, J. M. Chen, V. X. Shi, M. H. Ahn, A. M. Aalami, A. S. Aledia, E. M. Abrams, P. R. Cohen, J. Adler-Milstein, C. R.

Liao, Evaluating large language models for drafting emergency department encounter summaries, PLOS Digit. Health 4(6) (2025) e0000899. doi:10.1371/journal.pdig.0000899.

[5] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.

[6] L. Bednarczyk, J. N. Kather, D. Truhn, T. Raupach, M. Balyasnikova, M. Götz, Scientific evidence for clinical text summarization using large language models: scoping review, J. Med. Internet Res. 27 (2025) e68998. doi:10.2196/68998.

[7] 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, Proc. ACM Hum.-Comput. Interact. 5(CSCW1) (2021) Article 188, 1–21. doi:10.1145/3449287.

[8] M. Abdelwanis, H. K. Alarafati, M. M. S. Tammam, M. C. E. Simsekler, Exploring the risks of automation bias in healthcare artificial intelligence applications: a Bowtie analysis, J. Saf. Sci. Resil. 5(4) (2024) 460–469. doi:10.1016/j.jnlssr.2024.06.001.

[9] M. Naiseh, D. Al-Thani, N. Jiang, R. Ali, How the different explanation classes impact trust calibration: the case of clinical decision support systems, Int. J. Hum.-Comput. Stud. 169 (2023) 102941. doi:10.1016/j.ijhcs.2022.102941.

[10] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, R. Ranganath, A review of challenges and opportunities in machine learning for health, AMIA Jt. Summits Transl. Sci. Proc. 2020 (2020) 191–200.

[11] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (2020) 310. doi:10.1186/s12911-020-01332-6.

[12] X. Liu, S. C. Rivera, D. Moher, M. J. Calvert, A. K. Denniston, A.-W. Chan, A. Darzi, C. Holmes, C. Yau, H. Ashrafian, J. J. Deeks, L. Ferrante di Ruffano, L. Faes, P. A. Keane, S. J. Vollmer, A. Y. Lee, A. Jonas, A. Esteva, A. L. Beam, et al., Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, Nat. Med. 26(9) (2020) 1364–1374. doi:10.1038/s41591-020-1034-x.

[13] S. C. Rivera, X. Liu, A.-W. Chan, A. K. Denniston, M. J. Calvert, A. Darzi, A. L. Beam, H. Ashrafian, A. J. H. Smith, L. S. Bokhari, M. McCulloch, C. Holmes, A. Y. Lee, L. J. Mellor, A. Tufail, A. C. Cameron, A. K. M. Denniston, et al., Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension, Nat. Med. 26(9) (2020) 1351–1363. doi:10.1038/s41591-020-1037-7.

[14] VirtualCare, MedGPT - medical GPT revolutionizing healthcare with ethical AI, 2026.

[15] N. M. Kahl, M. J. Frieden, Z. R. Pope, M. M. Millen, V. M. Tolia, T. C. Chan, C. A. Longhurst, K. Singh, A. X. You, Evaluation of electronic health record-integrated artificial intelligence chart review, npj Health Syst. 3 (2026) 6. doi:10.1038/s44401-025-00064-x.

[16] S. W. Bell, O. de Rougemont, P. Heriot, T. D. Skinner, P. W. Burton, T. Louridas, T. V. Taylor, The National Aeronautics and Space Administration task load index (NASA-TLX): evaluation of its use in surgery, Surg. Endosc. 36 (2022) 940–947. doi: 10.1111/ans.17830.

[17] W. Armero, K. J. Gray, K. G. Fields, N. M. Cole, D. W. Bates, V. P. Kovacheva, A survey of pregnant patients' perspectives on the implementation of artificial intelligence in clinical care, J. Am. Med. Inform. Assoc. 30(1) (2023) 46–53. doi:10.1093/jamia/ocac200.

[18] R. Dlugatch, A. Georgieva, A. Kerasidou, AI-driven decision support systems and epistemic reliance: a qualitative study on obstetricians' and midwives' perspectives on integrating AI-driven CTG into clinical decision making, BMC Med. Ethics 25 (2024) 6. doi:10.1186/s12910-023-00990-1.