

Bridging the human-centered gap in agentic XAI: A multi-modal explanation user interface framework

Jason Wyatt, *Purdue University*, wyatt5@pnw.edu

George Stefanek, *Purdue University*, stefanek@pnw.edu

Abstract

As AI in healthcare advances toward agentic systems capable of multi-step reasoning and tool use, the primary barrier to clinical adoption is increasingly shifting from predictive accuracy to interpretability, reliability, and human-centered usability. This paper presents a human-centered Explanation User Interface (XUI) framework that combines reusable clinical interaction design patterns with a modular multi-agent architecture for structured explainable AI (XAI). To evaluate the framework, we introduce a reproducible benchmark that independently quantifies benchmark performance, numerical repeatability, and semantic stability using repeated production API executions across four representative BraTS2021 cases spanning the complete confusion matrix. The benchmark comprises 60 responses (4 benchmark subjects $\times$ 3 frontier LLMs $\times$ 5 repeated executions) evaluated using deterministic scoring, within-subject variability analysis, and embedding-based semantic similarity. Across repeated identical inputs, all evaluated models demonstrated low numerical variability and high semantic stability, although aggregate support. benchmark performance differed among models. By combining human-centered XUI design with reproducible computational benchmarking, this work provides a practical methodology for evaluating the reliability of structured LLM-generated SHAP interpretability reports and establishes a transferable framework for trustworthy human-AI collaboration in clinical decision.

Keywords: managing cognitive load, progressive disclosure, metadata-on-demand, patient-with-reference visualizations, adaptive feature attribution, delayed prediction

Introduction

The promise of generative AI is increasingly tempered by the opacity of large language models and their agentic implementations in the realm of high-stakes clinical environments. While diagnostic accuracy has advanced rapidly, a persistent socio-technical gap remains between model capabilities and clinical utility. As AI systems evolve toward multi-step reasoning and tool use, the complexity and stochasticity of their outputs often overwhelm clinicians, resulting in elevated cognitive load, inconsistent trust calibration, and explanation drift across repeated interactions.

This research addresses these challenges by developing a Human-Centered Explanation User Interface (XUI) for a brain MRI MGMT methylation classification use case. Building upon the reusable design patterns identified by Sheng et al. (2026), the systematic synthesis of XUI principles by Cappuccio et al. (2025), the influential HCI theories reviewed by Mohammadzadeh et al. (2025), and the human-centered

evaluation insights from Gambetti et al. (2026), we implement a modular architecture that employs dedicated API agents for each XAI modality. These specialized agents, guided by carefully engineered multi-modal prompts, dynamically generate tailored explanations in response to user requests and clinical context.

Due to space constraints, this paper focuses primarily on SHAP-based explanations while demonstrating a generalizable system designed to support multiple complementary modalities. To empirically ground our design, we examine intra-model consistency and inter-model variability across three leading frontier models by analyzing five sample responses per model to the same prompt, including representative cases from each quadrant of the patient confusion matrix.

By integrating proven design patterns such as delayed prediction, progressive disclosure, metadata-on-demand, adaptive feature attribution, patient-with-reference visualizations, and bias mitigation, the resulting XUI minimizes cognitive overload, better calibrates trust, and aligns with clinical workflows. Ultimately, this work offers a scalable, transferable framework for trustworthy, reproducible human-AI collaboration that extends beyond healthcare to other strategic decision-making domains.

Background and Literature Review

Human-AI interactions in clinical settings are systematically reviewed by Sheng et al. (2026) in *Design Patterns of Human-AI Interfaces in Healthcare* to establish a practical foundation for designing meaningful and efficient interfaces. In this work, the authors synthesize 12 reusable design patterns, and 15 information entities based on interviews with 12 healthcare professionals and a review of 43 research papers. Among the identified design patterns, four are particularly relevant to this work:

- **Delayed Prediction:** hides AI suggestions until a clinician completes an independent assessment to reduce confirmation bias.
- **Hiding Explanations by Default:** minimizes cognitive overload by revealing complex AI explanations, such as saliency maps, only when requested.
- **Interactive Cohort Selection:** enables clinicians to dynamically refine patient groups for iterative exploration.
- **Metadata on Demand:** surfaces implementation details, such as model provenance or architecture, only through optional interface interactions.

Workshop evaluations further demonstrated that these patterns help designers simplify complex interface structures while grounding interface decisions in authentic clinical user needs, highlighting their broader strategic value for developing more usable and trustworthy healthcare AI systems (Sheng et al., 2026). These design patterns extend beyond clinical AI and healthcare to offer a scalable, transferable framework for human-AI collaboration in any strategic decision-making environment. By prioritizing user trust, transparency, cognitive efficiency, interpretability, and human oversight, they enable trustworthy, Human-Centered Interfaces (HCI) that enhance organizational outcomes and decision autonomy.

A comprehensive overview of Explanation User Interfaces (XUIs) is given by Cappuccio et al. (2025) in the paper *Explanation User Interfaces: A Systematic Literature Review*. The visual and interactive systems that mediate between complex AI algorithms and human users are synthesized from insights gleaned by an examination of 146 academic papers to bridge the gap between technical XAI development and human-centered design. The urgent need for human-centered XAI is underscored in response to the "opaque" AI model problem, especially in deep learning systems, where decisions are produced that users cannot explain, raising concerns around bias, fairness, and trust. Cappuccio et al. define Explanation User Interfaces (XUIs) as the complete set of outputs and interactive elements that enable users to explore and

comprehend an AI's behavior. They advocate a two-stage design process that integrates technical model development with deliberate user interface design. The design of an Explanation User Interface (XUI) for a healthcare application, built upon generalizable design patterns, constitutes a primary focus of this work. This approach aligns with the broader Human-Centered AI (HCAI) framework that prioritizes augmenting human intelligence through greater transparency, accountability, and user empowerment.

The authors systematically classify core XAI techniques and organize the field around four central research questions. They distinguish between global explanations, which describe overall model behavior, and local explanations, which clarify individual predictions, while detailing prominent methods such as feature importance techniques (e.g., LIME and SHAP), saliency-based visualizations, prototype-driven examples, and counterfactual explanations. Their analysis further examines how task, domain, and user characteristics should shape XUI design (RQ1), the technical tools and models currently in use (RQ2), rigorous evaluation methodologies grounded in user studies and HCI principles (RQ3), and emerging design guidelines for practitioners (RQ4) (Cappuccio et al., 2025). The XUI developed in this work aligns closely with these proposed methodologies through the integration of multiple complementary explanation modalities, including SHAP, GradCAM, and Occlusion enabling both localized attribution analysis and interactive exploration of model behavior. By combining visual saliency methods with feature and perturbation attribution techniques, the interface supports more transparent and adaptive human-AI interaction while accommodating varying levels of user expertise and interpretability needs.

Nima Mohammadzadeh, Fatemeh Lotfi, and Hamed Samadpour (2025) provide a comprehensive analysis of the theoretical foundations shaping Health Information Technology (HIT) in their systematic review *Identifying Influential Theories in Human-Computer Interaction Within Health Informatics*. Through examination of 67 studies published between 1998 and 2024, the authors identify the most influential HCI theories and categorize them into three primary groups: technology acceptance models (e.g., TAM, UTAUT, TPB), behavior change theories (e.g., SCT, SDT), and sociotechnical and cognitive integration frameworks (e.g., Activity Theory, Normalization Process Theory). These frameworks have been applied across diverse systems including electronic health records (EHRs), clinical decision support systems (CDSS), mHealth applications, telehealth platforms, and hospital information systems, with UTAUT emerging as the most frequently used model, particularly in mHealth contexts.

The findings are especially relevant to explainable AI and clinical Explanation User Interface (XUI) development. The review underscores the critical role of user-centered design, cognitive load management, and sociotechnical alignment in driving usability, adoption, and long-term effectiveness. These insights directly support XUI design principles such as progressive disclosure, metadata-on-demand, delayed prediction, adaptive feature attribution (e.g., via SHAP, Grad-CAM, and occlusion), patient-with-reference visualizations, and bias mitigation. By grounding interfaces in established HCI theories, XUIs can better calibrate trust, enhance cognitive efficiency and task performance, and reduce mental overload for clinicians, while aligning with varying levels of digital literacy and clinical workflows. Ultimately, the paper reinforces that effective AI-assisted healthcare systems must integrate technical functionality with sociotechnical awareness, cognitive workload management, and trust-centered interaction design.

The current state of Human-Centered Evaluation (HCE) for AI in medical contexts is evaluated by Gambetti et al. (2026) in *A Survey on Human-Centered Evaluation of Explainable AI Methods in Clinical Decision Support Systems*. Using a PRISMA-guided systematic review of 31 human-centered evaluations of XAI in CDSS, the authors document a persistent gap between the rapid development of XAI algorithms and their limited real-world clinical effectiveness. Over 80% of the reviewed studies rely on post-hoc, model-agnostic techniques, most frequently SHAP and Grad-CAM, with evaluations typically conducted on small clinician samples (often fewer than 25 participants). While explanations are found to generally improve

trust and diagnostic confidence, they frequently increase cognitive load and show misalignment with clinicians' domain reasoning processes.

These findings strongly reinforce the relevance of the present work. Complementary research further highlights the challenges of reliable agent coordination and consistency. Berdoz et al. (2026) show that groups of LLM agents frequently fail to reach valid consensus even in benign, no-stake settings, with liveness failures (timeouts and stalled convergence) dominating, especially as group size grows or when even a single Byzantine agent (i.e., an unreliable or adversarial agent producing inconsistent behavior) is introduced. Rabanser et al. (2026) demonstrate that capability gains in modern agents have substantially outpaced improvements in reliability across consistency, robustness, predictability, and safety. In this work, we examine intra-model consistency and inter-model variability across three leading frontier models by analyzing five sample AI responses per model to the same prompt across representative cases from each quadrant of the patient confusion matrix.

To address the identified limitations, this research develops an Explanation User Interface (XUI) for a brain MRI MGMT methylation classification healthcare application that employs dedicated API agents for each XAI modality. These specialized agents dynamically generate tailored XAI translations in response to user requests and clinical context, enabling seamless, on-demand access to SHAP, Grad-CAM, and occlusion sensitivity within a single unified interface. Due to space constraints, this paper focuses primarily on SHAP-based XAI translation, while the underlying architecture is designed to support all modalities. The system is built upon generalizable design patterns including progressive disclosure, metadata-on-demand, delayed prediction, adaptive feature attribution, patient-with-reference visualizations, and bias mitigation.

By structuring agent interactions around well-defined, modality-specific roles guided by carefully engineered prompts, the architecture mitigates liveness and coordination failures while advancing multiple dimensions of reliability. Grounded in the HCI theoretical frameworks reviewed by Mohammadzadeh et al. (2025) and the stakeholder-centric, socio-technical approach advocated by Gambetti et al. (2026), the resulting XUI minimizes cognitive overload, better calibrates trust, aligns with clinical workflows, and enhances diagnostic accuracy while remaining transferable to other high-stakes human-AI decision-making environments.

Research Questions

This work is guided by the overarching research question: How can a multi-agent Explanation User Interface (XUI) architecture improve the usability, trustworthiness, and reliability of explainable AI in high-stakes clinical decision support? To address this question, the research investigates the extent to which established clinical HCI design patterns including delayed prediction, progressive disclosure, metadata-on-demand, adaptive feature attribution, and patient-with-reference visualizations can reduce cognitive overload and improve trust calibration when integrated into a multi-agent XUI framework. The study further examines how modality-specific explanation agents, guided through structured multimodal prompting strategies, may address coordination challenges, liveness concerns, and stochastic instability commonly observed in generalized large language model agent systems.

A secondary inquiry examines to what extent frontier large language models exhibit reproducible benchmark performance, numerical repeatability, semantic stability, and inter-model variability when generating structured explainable AI reports from multimodal clinical data. In response, this work evaluates benchmark performance, numerical repeatability, semantic stability, and inter-model variability across frontier large language models when generating XAI translations from identical multimodal clinical prompts containing both tabular and medical imaging data. Particular attention is given to how these

benchmark characteristics manifest across representative cases spanning all four confusion-matrix outcome categories using repeated production API executions. While the implementation presented in this work primarily emphasizes SHAP, the research additionally evaluates the broader generalizability of the proposed architecture across complementary XAI modalities including Grad-CAM, and occlusion sensitivity.

Finally, this work investigates the broader question: How can HCI-informed XUI design frameworks support trustworthy reproducible human-AI collaboration across healthcare and other high-stakes decision-making domains? To address this question, the study examines how HCI theories related to technology acceptance, cognitive load management, and sociotechnical systems can inform the design of clinically aligned XUI agent systems that better support clinician workflows, varying levels of digital literacy, and real-world decision-making constraints. Collectively, these research questions contribute toward the development of transferable frameworks for trustworthy, interpretable, and human-centered AI collaboration.

Methodology

This study employs a systems-design and reproducible benchmarking methodology to investigate how human-centered Explanation User Interface (XUI) design patterns can be integrated into a multi-agent architecture for explainable AI (XAI) in clinical decision support. Rather than performing clinical validation of diagnostic performance, the study focuses on the design, implementation, and evaluation of a prototype XUI framework through a reproducible benchmark that quantifies benchmark performance, numerical repeatability, semantic stability, and inter-model variability under repeated production API executions.

The methodology consists of two complementary components: development of a modular multi-agent XUI architecture grounded in established HCI and HCXAI design principles, and a deterministic benchmarking framework that evaluates structured LLM-generated XAI translation reports using repeated identical multimodal inputs, automated benchmark scoring, embedding-based semantic analysis, and provenance-tracked artifact generation.

Multi-Agent XUI Architecture

A modular XUI prototype was developed using dedicated modality-specific explanation agents responsible for generating structured XAI translation reports for complementary XAI techniques. Standardized multimodal prompts combined patient metadata, quantitative SHAP-derived metrics, and MRI segmentation/SHAP overlay images to provide consistent inputs across repeated benchmark executions. Although the architecture is designed to support multiple XAI modalities—including SHAP, Grad-CAM, and occlusion sensitivity—the present implementation focuses on SHAP because of its widespread adoption in HCXAI literature and its suitability for evaluating structured multimodal explanation generation and adhering to space requirements.

Human-Computer Interaction

The interface design incorporates established clinical HCI design patterns identified in prior HCXAI and healthcare interface literature (Sheng et al., 2026). These patterns were implemented to support cognitively manageable and workflow-aware interaction with AI-generated explanation:

- **Delayed Prediction:** AI-generated interpretations are withheld until an independent assessment is completed to reduce confirmation bias.

- **Progressive Disclosure:** High-level summaries are initially hidden along with detailed attribution maps, overlays, and cohort comparisons and are only revealed incrementally as requested by the user.
- **Metadata-on-Demand:** Technical information such as model provenance, confidence values, and methodological limitations remains accessible without dominating the primary interface.
- **Patient-with-Reference Visualizations:** Individual patient explanations are presented alongside cohort-level contextual references to support comparative interpretation.
- **Adaptive Feature Attribution:** Multiple complementary explanation modalities are incorporated to reduce overreliance on any single attribution method and support broader interpretability perspectives.

Human-Centered Evaluation Constructs

The benchmark design was informed by human-centered evaluation constructs derived from the HCXAI, HCI, and sociotechnical systems literature. These constructs guided the design and interpretation of the benchmark but were not independently quantified. Instead, empirical evaluation was performed using benchmark performance, numerical repeatability, and semantic stability, as described in the following sections.

- **Trust Calibration:** The extent to which explanation structure and uncertainty communication support appropriate user confidence in AI-assisted clinical decision-making.
- **Cognitive Efficiency:** The degree to which explanations facilitate rapid extraction of clinically relevant information while minimizing unnecessary cognitive burden.
- **Clinical Alignment:** The extent to which generated explanations present clinically interpretable reasoning and medically relevant contextual information.
- **Response Stability:** The consistency of generated explanations across repeated executions, operationalized in this benchmark through measures of numerical repeatability and semantic stability.

Comparative Benchmark Evaluation

To evaluate the reliability of structured LLM-generated SHAP XAI translation reports within the proposed XUI framework, a reproducible benchmark was conducted across three frontier LLMs. Identical multimodal production API requests containing patient metadata, quantitative SHAP-derived metrics, and MRI segmentation/SHAP overlay images were submitted five independent times to each model. The benchmark independently quantified three complementary evaluation dimensions.

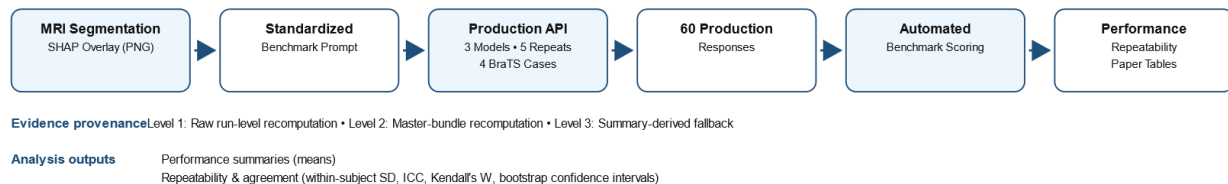


Figure 1. Reproducible Benchmark Evaluation Pipeline

Figure 1 summarizes the reproducible benchmark evaluation pipeline, from standardized multimodal inputs through production API execution, deterministic scoring, semantic analysis, and provenance-tracked artifact generation.

- **Benchmark Performance:** Deterministic benchmark scoring measured overall criterion satisfaction (Present Rate) and unmet required benchmark elements (Required-Risk Rate).
- **Numerical Repeatability:** Within-subject standard deviations quantified the stability of benchmark scores across five repeated executions for each subject-model pair.

- **Semantic Stability:** Mean pairwise cosine similarity of sentence embeddings quantified the semantic consistency of repeated responses independent of lexical variation.

To evaluate inter-model variability, the same benchmark protocol was applied to gpt-4.1-mini, grok-3-mini, and gemini-3.5-flash. Four representative BraTS2021 subjects, spanning the complete confusion matrix (true negative, true positive, false negative, and false positive), were evaluated using identical multimodal inputs, yielding a total of 60 benchmark responses (4 subjects $\times$ 3 models $\times$ 5 repeated executions). All responses were obtained through authenticated production API calls and processed through a deterministic, provenance-tracked evaluation pipeline to ensure reproducibility.

Results

In high-stakes clinical environments, correctness alone is insufficient. Clinicians require AI systems that deliver accurate insights in ways that are cognitively efficient, trustworthy, and seamlessly integrated into their workflows. To address this need, we implemented and evaluated a multi-agent Explanation User Interface (XUI) prototype that operationalizes several established human-centered design patterns. The following results demonstrate how these patterns manifest in practice, supported by representative screenshots from the live interface.

The available patients with data present are hidden until the button is clicked, at which point all the patients' ids are displayed in a neat tabular format. The patient id can be chosen by clicking inside the appropriate table cell and the id is filled forward as the default patient id in all subsequently labeled textboxes for convenience, and to eliminate a mismatch by a possibly overlooked keyboard input typo. The table can then be collapsed to reduce visual clutter as seen in Figure 2. The MGMT status and confusion matrix bucket information are not visible until the user toggles the button. Also, the chosen subject remains visible as the active sample when the table is collapsed.

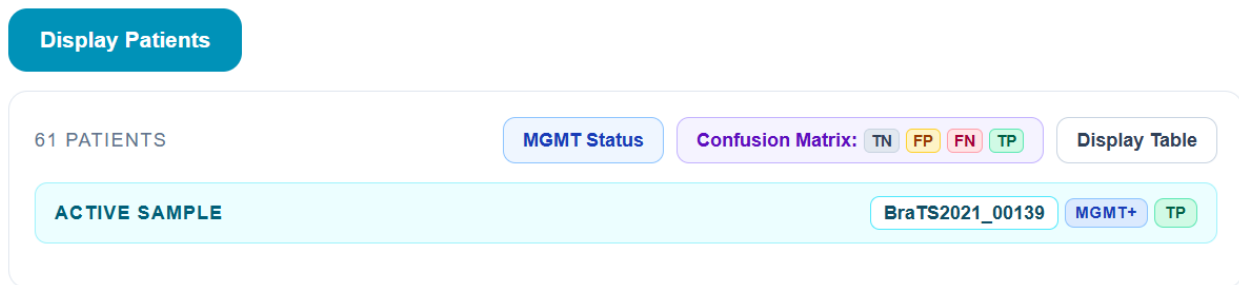


Figure 2. SHAP Analysis Interactive Cohort Selection with MGMT Status and CM Bucket

A core principle guiding the interface is *delayed prediction with progressive disclosure* (Sheng et al., 2026). High-level summaries, key predictions, and detailed explanations remain hidden by default and are revealed only upon explicit user request. This approach prevents information overload and mitigates confirmation bias during routine cases while preserving access to deeper reasoning when needed.

Figure 3 illustrates the patient-level MRI inspection interface designed to support clinician-driven exploration before AI-generated explanations are introduced. The interface enables simultaneous visualization of up to three MRI modalities, allowing clinicians to compare complementary imaging perspectives within a coordinated visual space. Prior to displaying explanatory overlays or feature attributions, users may independently inspect image slices, enlarge visualizations to full-screen mode, and perform localized magnification through interactive mouse controls. This design supports progressive disclosure and delayed prediction by preserving the clinician's initial diagnostic reasoning process before

exposing model-generated interpretability information. By separating raw image inspection from explanatory augmentation, the interface mitigates bias effects while supporting cognitive focus and contextual understanding across modalities.

MRI Slice Viewer

Enter patient ID, choose slice and modality, then load the image slice.

Patient ID

Slice (0-155)

BraTS2021_00139

66

Modalities (up to 3)

FLAIR

T1CE

T1

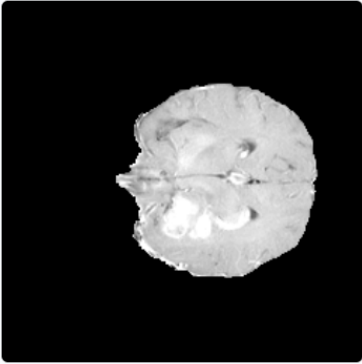
T2

SEG

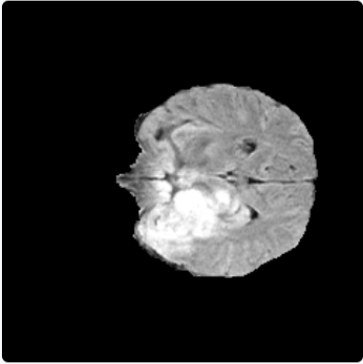
Selected: T1CE, FLAIR, T2

Load Slice

T1CE



FLAIR



T2

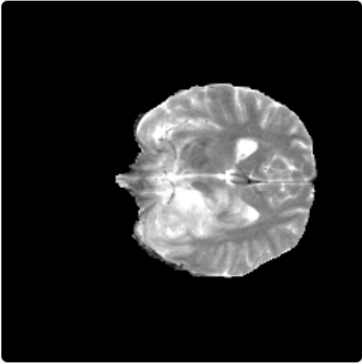


Figure 3. Patient-Level Multi-Modal MRI Inspection Interface with Synchronous Multimodal Slice Display Prior to Explanation Overlay

Figure 4 presents the instance-level SHAP summary for a selected patient. The table lists key prediction and explanation metrics, including predicted probability, tumor composition, and top contributing features. Clinicians can flexibly select the information they wish to interrogate through interactive checkboxes. A right click drag enables rapid multi-row (un)selection, streamlining explanation curation and reducing interaction overhead. Only the selected metrics are incorporated into the API prompt for downstream explanation generation and disclosure. The global SHAP summary interface (which has been omitted for space) has the same interface functionality, and it aggregates explanation metrics across outcome categories (FN, FP, TN, TP). This view supports cohort-level interpretation by summarizing tumor composition statistics and feature attribution patterns across patient groups.

These UI interface features enhance flexibility by supporting adaptive feature attribution, enabling users to manage the cognitive load generated in the AI explanation. Finally, this delayed assessment mechanism withholds AI-generated recommendations until the clinician has submitted a customized feature assessment

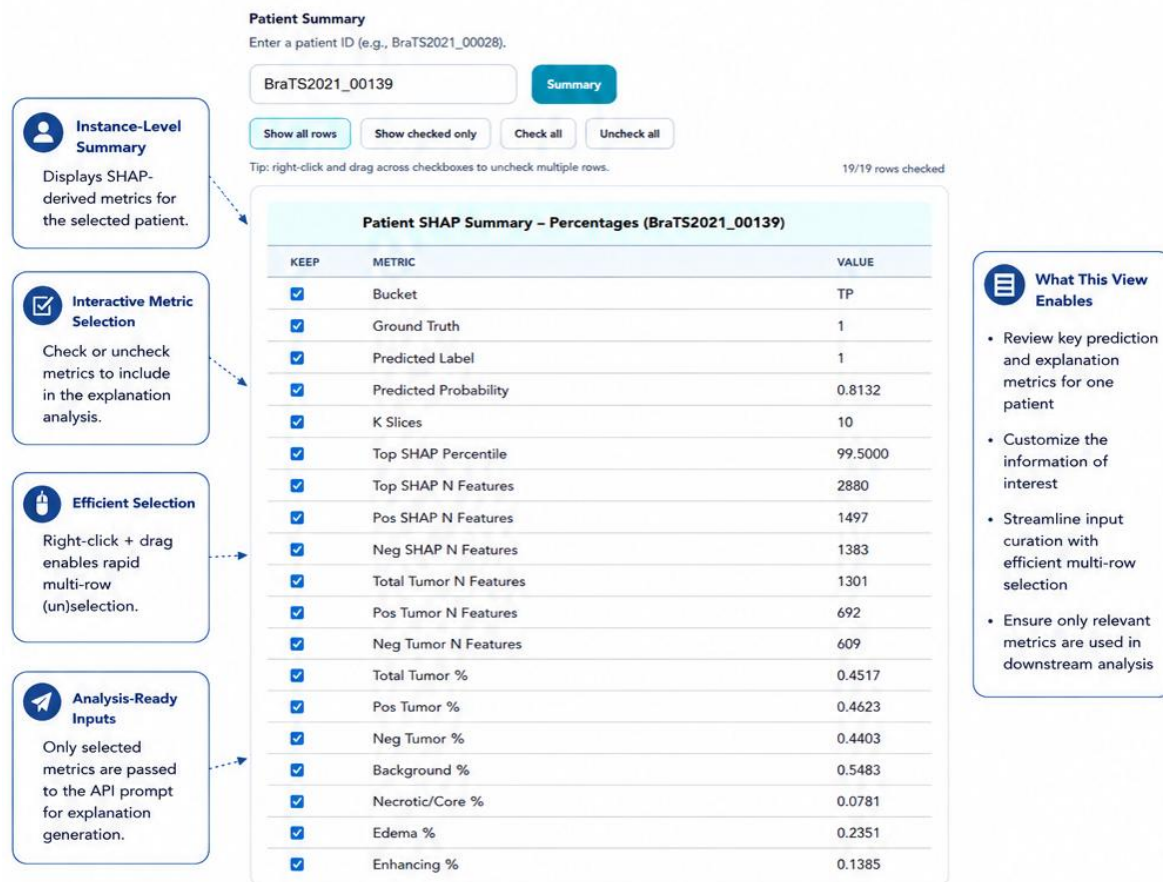


Figure 4. Instance Level Drill-Down SHAP Summary with Interactive Metric Selection

prompt request, further mitigating confirmation bias and promoting a more balanced human-AI collaboration. Similarly, patient-with-reference visualizations allow clinicians to rapidly contextualize individual patient data against relevant reference ranges and population trends within the same view, reducing the need for context switching. These figures are omitted to include the intra-model consistency and inter-model variability results.

The SHAP visualization panel with a custom toolbar for segmentation and top-k (%) SHAP features (pixels) overlay for the model inference can be seen in Figure 5. The SRI24 anatomical mapping can be overlaid on the sample because they have the same space registration (there is some evident drift). This mapping makes statistical analysis of what anatomical regions the tumor occupies by percentage and label possible. Also, the alpha value for the segmentation label overlays can be adjusted.

The ‘Compose API Viz’ button creates a .png file from the current visualized image that will be shared with the model to be included in its response. Because the visualization is central to how humans interpret the model, the language model should analyze the exact same visual artifact as the reviewer. This shared context allows the model to ground its explanations in direct graphical evidence rather than raw prediction outputs, making the explanations more faithful, interpretable, and aligned with the user's perspective.

The XUI offers mouse wheel control zoom +/-, plus a fit and reset function, and a full screen viewing mode with zooming capabilities and image movement mouse control. From full screen mode a close button is offered. One MRI scan image with toolbar buttons offering concurrent separate overlays of segmentation, heatmap, and top SHAP features reduces clutter, improves focus, and offers better visual cuing.

SHAP Viz

Generate and display the SHAP visualization image for a patient. Slice options are loaded automatically for the entered patient ID.

SHAP Viz

Loaded 10 slice options for BraTS2021_00139.

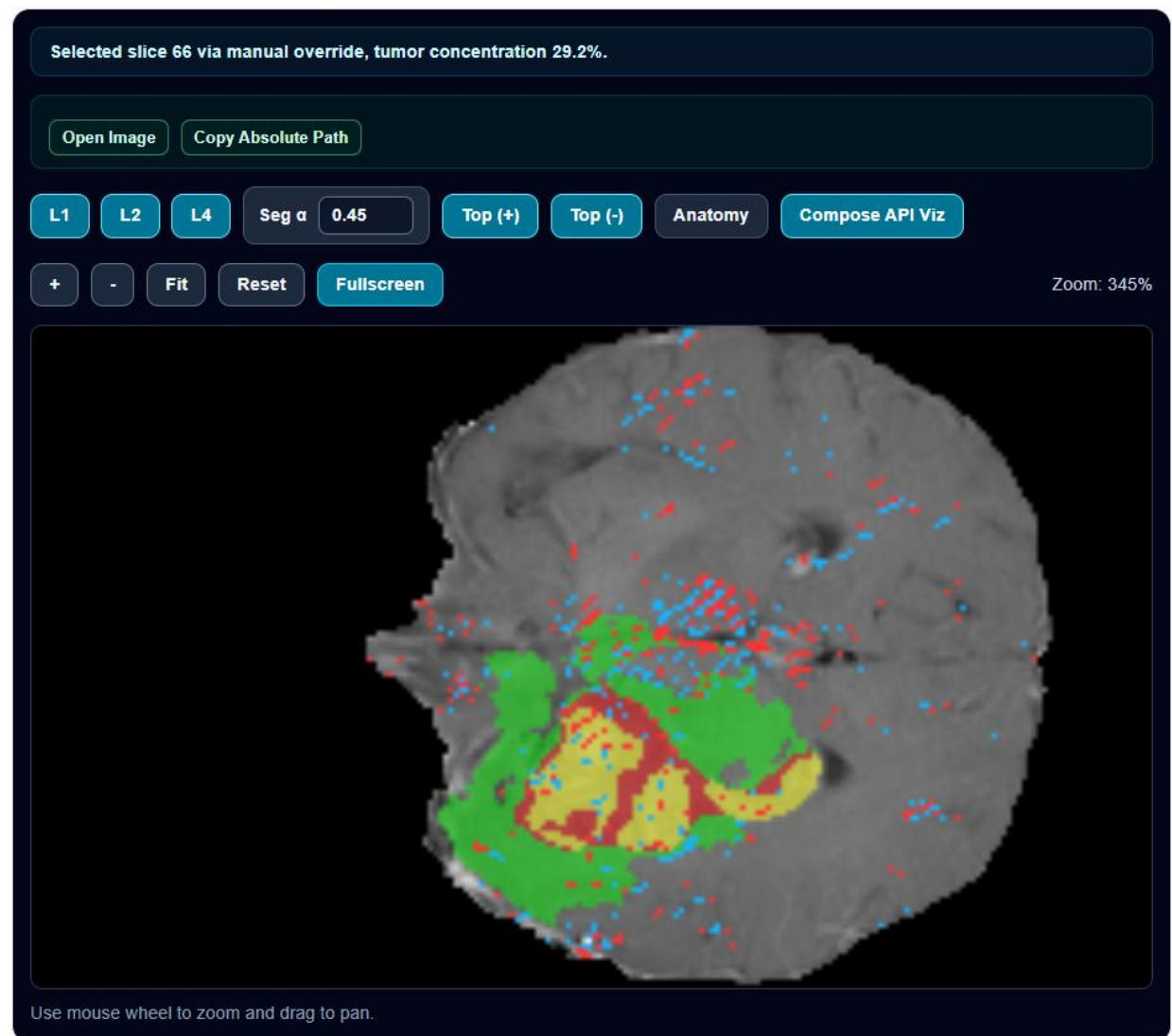


Figure 5. Segmentation and SHAP Top-Feature % (+/-) Overlay

The XUI dashboard, omitted for space, provides a Multi-agent XAI translation panel that supports a structured pipeline: provider selection, data attachment, API prompt generation, and API invocation. As these cascading actions occur, downstream interface sections become visible in order, reflecting the current state of the translation workflow. While waiting for the response as an API Call is made, the status of the

https request and responses become visible. Also, when the API response is received, it becomes visible in the ‘Model SHAP Translation’ textbox that appears.

Together, these patterns prioritize cognitive efficiency and clinical alignment by exposing system state in real time, minimizing hidden workflow steps, and presenting the final SHAP translation exactly where decisions are made. This progressive disclosure model supports safer interpretation workflows by pairing operational transparency (request and response status) with immediate access to the final model-generated SHAP translation in context.

Discussion of Findings

We evaluated the repeatability of LLM-generated SHAP interpretability reports across four BraTS2021 benchmark cases representing the complete confusion matrix (TN: 00347, TP: 00504, FN: 00138, and FP: 01009). Sixty production API responses (4 subjects $\times$ 3 models $\times$ 5 repeated executions) were analyzed using a reproducible benchmark that independently quantified three complementary dimensions of reporting behavior: benchmark performance, numerical repeatability, and semantic stability. Benchmark performance was summarized using mean Present Rate and Required-Risk Rate, numerical repeatability was quantified using within-subject standard deviations across repeated executions, and semantic stability was quantified using mean pairwise cosine similarity of sentence embeddings. The following results summarize these benchmark characteristics across all evaluated models.

Semantic Stability Assessment

Semantic stability was quantified computationally rather than through manual qualitative review. For each subject–model cell, the five repeated responses were encoded using the fixed sentence-transformers/all-MiniLM-L6-v2 sentence embedding model, and pairwise cosine similarity was computed for all ten unique response pairs ($\binom{5}{2} = 10$). The mean within-subject cosine similarity was used as the primary measure of semantic stability, with higher values indicating greater semantic consistency across repeated responses. Semantic dispersion was defined as $1 - \text{mean cosine similarity}$, where lower values indicate less semantic variability.

To ensure reproducibility, semantic analysis was performed only when all five responses were successfully recovered for a subject–model cell. Cells with fewer than five responses were classified as *not estimable* and excluded from aggregate semantic stability summaries.

Benchmark Results

The benchmark successfully completed all planned evaluations, producing a complete dataset of 60 production API responses (4 benchmark subjects $\times$ 3 models $\times$ 5 repeated executions). All subject–model cells satisfied the predefined benchmark design, allowing computation of benchmark performance, numerical repeatability, and semantic stability for each evaluated model.

As summarized in Table 1, GPT-4.1-mini achieved the highest aggregate benchmark performance, with the greatest mean Present Rate (49.04%) and the lowest mean Required-Risk Rate (18.95%). Gemini 3.5 Flash (44.30% and 21.23%) and Grok 3 Mini (43.95% and 22.02%) demonstrated comparable aggregate benchmark performance. Across all evaluated models, numerical repeatability remained high, with low within-subject standard deviations observed across repeated API executions. Mean within-subject Present

Rate standard deviations ranged from 1.39 to 2.65 percentage points, while Required-Risk Rate standard deviations ranged from 0.44 to 1.18 percentage points.

Table 1. Summary of Benchmark Performance, Repeatability, and Semantic Stability

Metric	GPT-4.1-mini	Gemini 3.5 Flash	Grok 3 Mini	Interpretation
Mean Present Rate (%)	49.04	44.30	43.95	Higher indicates greater benchmark satisfaction.
Mean Required-Risk Rate (%)	18.95	21.23	22.02	Lower indicates fewer unmet required benchmark elements.
Mean Within-Subject SD (Present Rate)	1.94	1.39	2.65	Lower indicates greater numerical repeatability.
Mean Within-Subject SD (Required-Risk Rate)	0.44	1.18	0.92	Lower indicates greater numerical repeatability for required benchmark elements.
Mean Within-Subject Cosine Similarity	0.9068	0.8635	0.8905	Higher indicates greater semantic stability across repeated responses.
Semantic Dispersion	0.0932	0.1365	0.1095	Lower indicates less semantic variability (1 – cosine similarity).

Embedding-based semantic analysis likewise demonstrated high semantic stability across repeated responses. Mean within-subject cosine similarity ranged from 0.8635 to 0.9068, corresponding to low semantic dispersion (0.0932–0.1365). These findings indicate that repeated responses consistently conveyed similar semantic content despite expected differences in wording and sentence structure.

Table 2. Aggregate Benchmark Performance and Numerical Repeatability

Model	Mean Present Rate (%)	Mean Required-Risk Rate (%)	Mean SD (Present)	Mean SD (Required Risk)	Pooled SD (Present)	Pooled SD (Required Risk)
GPT-4.1-mini	49.04	18.95	1.94	0.44	2.04	0.62
Gemini 3.5 Flash	44.30	21.23	1.39	1.18	1.52	1.24
Grok 3 Mini	43.95	22.02	2.65	0.92	2.76	1.07

Table 3. Aggregate Semantic Stability

Model	Mean Within-Subject Cosine Similarity	Semantic Dispersion
GPT-4.1-mini	0.9068	0.0932
Gemini 3.5 Flash	0.8635	0.1365
Grok 3 Mini	0.8905	0.1095

The detailed benchmark metrics are presented in Tables 2 and 3. Aggregate values were computed from benchmark results spanning all four benchmark subjects representing the complete confusion matrix (true negative, true positive, false negative, and false positive). Subject-level results are omitted because of space constraints but are available in the reproducible benchmark artifacts. Collectively, these findings indicate

that repeated production API executions produced numerically repeatable and semantically stable SHAP interpretability reports under the evaluated benchmark protocol.

Key Highlights & Pertinent Information

- GPT-4.1-mini achieved the highest aggregate benchmark performance under the evaluated protocol, demonstrating the greatest overall benchmark criterion satisfaction (highest Present Rate) while exhibiting the fewest unmet required benchmark elements (lowest Required-Risk Rate).
- All three models demonstrated low numerical variability across repeated executions, indicating that submitting identical inputs multiple times produced highly consistent benchmark scores rather than substantially different benchmark outcomes.
- High mean within-subject cosine similarity indicated strong semantic stability despite expected differences in wording, showing that repeated responses generally conveyed the same underlying interpretation even when phrasing or sentence structure varied.
- Aggregate benchmark metrics were derived from all four confusion-matrix categories (TN, TP, FN, and FP), reducing the likelihood that the reported results were dominated by a single representative case.
- By independently quantifying benchmark performance, numerical repeatability, and semantic stability, the benchmark provides complementary evidence for evaluating the reliability and reproducibility of structured LLM-generated SHAP interpretability reports, rather than relying on a single aggregate performance measure.

Implications

This study demonstrates that structured XAI translation reports can be evaluated using a reproducible benchmark that independently measures benchmark performance, numerical repeatability, and semantic stability. Across 60 production API responses spanning all four confusion-matrix categories, all evaluated models exhibited low numerical variability and high semantic stability, indicating that repeated submissions of identical inputs produced consistent benchmark outcomes and semantically similar explanations.

Although aggregate benchmark performance differed among the evaluated models, the consistently low within-subject variability suggests that standardized prompting and structured output criteria can produce highly repeatable interpretability reports. By separating benchmark performance from repeatability and semantic stability, the framework provides a more comprehensive assessment of reporting reliability than aggregate scores alone.

Beyond comparing individual models, the principal contribution of this work is the benchmark itself. The framework integrates deterministic scoring, embedding-based semantic analysis, reproducible statistics, and provenance-tracked artifact generation into a reproducible evaluation pipeline that supports independent verification and longitudinal assessment of structured medical XAI reporting.

Limitations

The benchmark evaluated four representative BraTS2021 cases (one per confusion-matrix category) with five repeated production API executions per model. Although appropriate for controlled repeatability assessment, the results may not generalize to larger datasets, other imaging modalities, or alternative prompting strategies.

The benchmark evaluates reproducibility rather than clinical correctness or explanation faithfulness. Similarly, sentence-embedding cosine similarity provides a computational estimate of semantic consistency, not definitive equivalence of clinical meaning.

Future Work

Future work will investigate clinician-centered validation, longitudinal benchmark monitoring, and support for complementary XAI techniques beyond SHAP, including Grad-CAM and occlusion sensitivity. Additional research will explore enhanced structured output strategies, including schema-constrained generation (e.g., JSON Schema and Pydantic), DSPy-style prompt optimization, and interface-level guardrails to further reduce response variability and strengthen adherence to standardized clinical interpretability templates.

Collectively, these efforts aim to extend the proposed benchmark from a controlled evaluation framework toward a standardized, scalable methodology for reproducible assessment of LLM-generated clinical XAI explanations.

Conclusion

This study presents a reproducible benchmark for evaluating structured LLM-generated XAI translation reports within a human-centered Explanation User Interface (XUI) framework for clinical decision support. Using repeated production API executions across four representative BraTS2021 cases, the benchmark independently quantified benchmark performance, numerical repeatability, and semantic stability through deterministic scoring, within-subject variability analysis, and embedding-based semantic similarity.

Across 60 benchmark responses, all evaluated models demonstrated low numerical variability and high semantic stability under repeated identical inputs, although aggregate benchmark performance differed among models. These findings suggest that structured LLM-generated interpretability reports can achieve a high degree of reproducibility when standardized prompting, deterministic evaluation, and structured reporting criteria are employed.

Beyond the comparison of individual models, the principal contribution of this work is the benchmark itself. By integrating deterministic benchmark scoring, semantic stability analysis, provenance-tracked artifact generation, and reproducible evaluation workflows, the framework provides a transparent methodology for independently assessing the reliability of structured medical XAI reporting.

Reproducible evaluation will become as important as model performance itself as LLM-assisted clinical decision support continues to mature. The proposed benchmark provides a foundation for future benchmarking efforts across larger clinical cohorts, additional XAI modalities, emerging frontier language models, and prospective clinician-centered validation.

Acknowledgements

Generative AI tools were used to assist with language refinement, editing, software development, and code review during the preparation of this manuscript. All study design, benchmark development, experimental methodology, data processing, analyses, interpretation of results, and manuscript revisions were performed, reviewed, and validated by the authors. The authors accept full responsibility for the accuracy, integrity, and originality of the work.

References

- Berdoz, F., Rugli, L., & Wattenhofer, R. (2026). *Can AI agents agree?*. arXiv. <https://arxiv.org/abs/2603.01213>.
- Cappuccio, E., Esposito, A., Greco, F., Desolda, G., Lanzilotti, R., & Rinzivillo, S. (2025). *Explanation user interfaces: A systematic literature review*. arXiv. <https://arxiv.org/abs/2505.20085>.
- Gambetti, A., Han, Q., Shen, H., & Soares, C. (2026). *A survey on human-centered evaluation of explainable AI methods in clinical decision support systems*. arXiv. <https://arxiv.org/abs/2502.09849>.
- Mohammadzadeh, N., Lotfi, F., & Samadpour, H. (2025). *Identifying influential theories in human-computer interaction within health informatics: A systematic review*. *Health Science Reports*, 8(7), e71074.
- Rabanser, S., Kapoor, S., Kirgis, P., Liu, K., Utpala, S., & Narayanan, A. (2026). *Towards a science of AI agent reliability*. arXiv. <https://arxiv.org/abs/2602.16666>
- Sheng, R., Shi, C., Lotfi, S., Liu, S., Perer, A., Qu, H., & Cheng, F. (2026). *Design patterns of human-AI interfaces in healthcare*. *International Journal of Human-Computer Studies*, 103737.