

Who Are You Explaining To?

A Multi-Agent System for Audience-Aware XAI Narratives

Francesco Musicco
Politecnico di Bari, Italy
francesco.musicco@poliba.it

Danilo Danese
Politecnico di Bari, Italy
danilo.danese@poliba.it

Giuseppe Fasano
Politecnico di Bari, Italy
giuseppe.fasano@poliba.it

Angela Lombardi
Politecnico di Bari, Italy
angela.lombardi@poliba.it

Alberto Carlo Maria Mancino
Politecnico di Bari, Italy
alberto.mancino@poliba.it

Tommaso Di Noia
Politecnico di Bari, Italy
tommaso.dinoia@poliba.it

Abstract

Feature-attribution methods such as SHAP provide useful evidence about individual model predictions, but their numerical outputs are rarely sufficient for audiences with different expertise, goals, and risks of misinterpretation. In medical AI, the same local explanation must reach patients, clinicians, and data scientists through markedly different forms of communication, and naive verbalization through large language models (LLMs) is prone to weak grounding, conflation of attribution with causal language, and outputs that are persuasive without being faithful to the underlying model evidence. We introduce XstrAI, an audience-aware multi-agent framework that treats local explanations as fixed evidence and structures how it is communicated to each target reader. Each prediction case is encoded as an immutable structured representation, shared identically across audiences so the underlying evidence remains fixed. Generation is factored into three specialized LLM agents responsible for audience-aware planning, linguistic realization, and validation for grounding, attribution consistency, communicative risk, and audience appropriateness, with a bounded revision loop triggered on detected inconsistencies. We evaluate XstrAI on diabetes and stroke risk prediction against 11 baselines, ranging from direct verbalization to a re-implementation of a state-of-the-art narrator. The evaluation combines an intra-narrative regime measuring fidelity to SHAP evidence with an extra-narrative regime assessing audience appropriateness through reference corpora, multi-family LLM judges, and a survey with target readers. In both evaluations, XstrAI's narratives are consistently assigned to their intended audience by independent judges, and preferred over all baselines on Clinician and Patient audiences, with competitive performance on Data Scientist, where audience-conditioned single-prompt baselines lead. We make the code publicly available at <https://github.com/sisinflab/XstrAI>.

CCS Concepts

- **Computing methodologies** → Natural language processing;
- **Information systems** → Collaborative and social computing systems and tools.

Keywords

XAI, Agent Based AI, LLM, SHAP, Narrative XAI

1 Introduction

Local feature-attribution methods such as SHAP have become a standard component of post-hoc interpretability pipelines [32], providing per-instance evidence about model behavior in the form of feature contributions, ranks, and directions. As these methods are deployed in operational settings, the audience of an explanation has emerged as a determinant of its usefulness no less important than its technical fidelity: the same attribution output supports different inferences, different admissible claims, and different downstream actions depending on whether it is consumed by a domain expert, a non-expert decision subject, or a model developer [10, 28, 36, 39]. Empirical work on human-AI collaboration further shows that explanations misaligned with the reader can induce inappropriate reliance and degrade joint performance, even when the underlying evidence is technically correct [1, 31]. Audience adaptation is therefore a condition for the safe and effective use of XAI outputs, rather than a presentational refinement.

Existing approaches address this requirement only partially. Feature-attribution methods are by design audience-agnostic, leaving the transformation of evidence into communicable content to whoever consumes the output. Direct verbalization of attribution values through large language models (LLMs), or single-prompt role conditioning, are the natural shortcut, but recent evidence indicates that such generation is prone to weak grounding [22], to a conflation of attribution with causal language [48], and to outputs that are persuasive without being faithful to the underlying model evidence [31, 43]. Narrative-XAI systems improve accessibility but largely treat generation as a monolithic step, in which evidence selection, audience adaptation, linguistic realization, and validation are entangled within a single pass and offer limited inspectable control [9, 33, 51].

In this work we introduce XstrAI, a multi-agent framework for audience-aware communication of local explanation evidence. Rather than treating explanation generation as a monolithic narration task, XstrAI structures explanation communication as a controlled transformation pipeline grounded in a shared, immutable representation of the prediction case. Each instance is encoded as a machine-readable artifact that fixes the prediction, attribution values, feature semantics, and generation constraints identically across stakeholder pathways. Within each pathway, generation is factored into three role-specific LLM agents responsible for stakeholder-aware planning, linguistic realization, and validation against grounding, attribution consistency, communicative risk, and audience appropriateness, with a bounded revision loop triggered



This work is licensed under a Creative Commons Attribution 4.0 International License.

when inconsistencies are detected. We instantiate the framework on two clinical prediction tasks, diabetes and stroke risk estimation, where stakeholder heterogeneity is well documented and the cost of miscommunication is concrete. We evaluate XstrAI against 11 baseline configurations, including a re-implementation of Explingo [51] and a family of incremental ablations. The evaluation is organized around two complementary regimes that reflect the dual nature of the problem: an *intra-narrative* regime measuring text form, audience-specific semantics and fidelity of the produced narrative to the underlying explanation evidence, and an *extra-narrative* regime measuring its appropriateness for the target reader against audience-specific human reference corpora, three independent LLM judge families aggregated via a Bradley-Terry/Elo model, and a human study with representatives of each target audience. Our contributions are threefold:

- A novel architecture for audience-aware XAI narrative generation that decouples evidence representation, stakeholder-specific planning, linguistic realization, and validation into inspectable stages.
- A dual evaluation protocol that disentangles intra-narrative fidelity to the underlying explanation evidence from extra-narrative appropriateness for the target audience, with the latter assessed against audience-specific human reference corpora, multi-family LLM judges, and a human study.
- An empirical study on diabetes and stroke prediction against 11 baseline configurations, with rank-based statistical analysis of the audience-specific outcomes.

2 Related Work

2.1 XAI and Stakeholder Needs

Local explanation methods occupy a central role in the XAI literature by providing instance-level evidence for individual model predictions. Among them, feature-attribution methods such as SHAP [32] estimate the contribution of input features to a prediction, enabling users to inspect and interpret the factors influencing the model output at the level of individual instances. These methods are widely adopted for model debugging, feature attribution analysis, and post-hoc inspection of black-box systems.

The increasing use of local explanations has progressively exposed a distinction between explanation evidence and explanation communication. Feature attribution scores, ranked feature lists, and visualization-based outputs provide technically informative representations of model behavior, but their interpretation depends on the reader’s expertise, objectives, and contextual knowledge [24, 31]. Prior work in human-centered XAI emphasizes that explanation quality cannot be reduced to technical properties such as fidelity or stability, since explanations operate as communicative artifacts shaped by human goals and decision-making contexts [28, 34], and empirical studies on human-AI collaboration further show that explanations misaligned with the reader can induce inappropriate reliance even when the underlying evidence is technically correct [1, 31]. This concern is particularly acute in multi-stakeholder environments, where the same predictive evidence must support different interpretive goals: clinicians typically require technically grounded inspection of feature contributions, patients require accessible and non-alarming communication, and data scientists focus

on attribution behavior and model dynamics [10, 30, 36, 39]. Existing local explanation methods provide valuable evidence for understanding individual predictions but offer limited support for controlled stakeholder-specific communication.

2.2 Natural Language and Narrative XAI

Recent work has explored natural language as a means to make explanation evidence more accessible and actionable with LLM-based approaches transforming model outputs and attribution artifacts into textual explanations [2]. Systems such as Explingo [51] focus on prediction-level explanation generation, narrative-driven XAI [33] frames explanations as structured stories, and Contextual-SHAP [9] enriches SHAP-based explanations through contextual language generation. More recent work has further emphasized personalization and trustworthiness in natural language explanations, reflecting the growing interest in explanation generation as a communication-oriented layer of XAI [44]. This direction marks an important shift from raw numerical or visual artifacts toward more interpretable textual forms, supporting contextualization, selective emphasis, and coherent presentation of local evidence.

Fluent explanation generation, however, introduces specific risks. Textual explanations may appear coherent while weakening the connection with the underlying evidence, particularly when produced directly from attribution values without explicit validation: direct generation by LLMs can introduce hallucinations and unsupported claims [22], and language models are known to produce plausible explanations that misrepresent the actual factors influencing a decision [43], with the additional risk of conflating attribution with causal inference [48]. These observations are consistent with the broader distinction between explanations that are persuasive to users and explanations that remain substantively aligned with the model evidence [31], and reinforce the view that explanation quality requires multiple dimensions beyond fluency [36].

Existing XAI narratives approaches primarily improve the accessibility of explanation evidence while leaving open the problem of controlled generation. In particular, current systems provide limited explicit separation between evidence selection, stakeholder-specific framing, linguistic realization, and validation, motivating the need for explanation generation pipelines in which narrative production is treated as a structured and verifiable process rather than as a direct transformation from explanation values to text.

2.3 Multi-Agent and Role-Based LLM Systems

Recent advances in LLMs have stimulated growing interest in multi-agent and role-based interaction frameworks for complex reasoning tasks, in which language models are organized into specialized roles that interact through collaboration, critique, or coordinated task decomposition [27, 38]. Parallel research on role-playing and persona prompting has shown that agent behavior can be guided through explicit role assignments and evaluated in terms of behavioral consistency and task-oriented effects [6, 47], while multi-agent debate frameworks demonstrate that structured interaction between language models can improve factual consistency and reduce unsupported claims by exposing disagreements and refining intermediate outputs [8]. Within domain-specific reasoning, MedAgents [42]

shows that collaborative interaction between multiple language-model agents can improve medical reasoning through distributed analysis and iterative discussion, a setting that resonates with the clinical instantiation considered in this work.

Current multi-agent systems focus on reasoning performance [8], collaborative problem solving [20], behavior simulation [38], or persona fidelity [45]. The communication of fixed XAI evidence introduces a different requirement structure, in which the central objective is the controlled transformation of explanation evidence into stakeholder-specific narratives while preserving grounding and attribution consistency. Existing approaches provide useful foundations for role specialization and iterative refinement, while offering limited support for explanation-specific validation and explicit control over how explanatory evidence is selected, adapted, and verified throughout the generation process. Our work bridges these research strands by treating local explanations as fixed evidence to be communicated, and by integrating narrative XAI, stakeholder adaptation, and multi-agent control within a single explanation-communication pipeline instantiated on SHAP attributions.

3 XstrAI Framework

Figure 1 illustrates **XstrAI**, the proposed framework for transforming local feature-attribution explanations into narratives tailored to different target audiences, or stakeholders, that we use interchangeably in the rest of the paper. Operating at the explanation-communication layer, the framework keeps the predictive output and attribution values fixed, while controlling how the same evidence is selected, organized, verbalized, and validated for each target reader.

Given the clinical nature of the considered scenarios, XstrAI is instantiated around three stakeholder profiles: clinician, patient, and data scientist. These profiles capture different explanatory needs, requiring the same local explanation to be transformed into audience-specific narratives from a shared evidential basis.

The architecture is built around an *ExplanationCard*, which provides the structured case-level evidence and serves as source of truth for the pipeline. From this representation, three agentic components operate within each stakeholder-specific pathway: the *Framer* plans the explanation, the *Narrator* generates the narrative, and the *Reviewer* validates the output through bounded revision cycles when needed. Intermediate and final artifacts are retained to support traceability and inspection. The modular design is motivated by recent multi-agent LLM research highlighting the benefits of role-based task decomposition and explicit critique mechanisms [8, 20].

3.1 Input Representation and Control Layer

Each prediction case is represented through an *ExplanationCard*, a machine-readable source of truth for the generation process. Inspired by the Model Cards paradigm [35], which standardises documentation at the *model* level, the *ExplanationCard* transposes the same structured-artifact idea to the *instance* level: every prediction case is paired with a self-contained, machine-readable specification of the evidence to be communicated. This specification includes case information, model output, SHAP attributions, feature metadata, generation constraints, and the stakeholder profile.

The *ExplanationCard* acts as a structured data contract between the predictive layer and the narrative generation pipeline: all stakeholder pathways receive the same validated representation, so differences across narratives depend on communicative choices rather than on changes in the underlying evidence. A deterministic controller orchestrates each pathway by passing the card to the *Framer*, forwarding the resulting *FramingTrace* to the *Narrator*, and sending the candidate narrative to the *Reviewer*, which either accepts the output or triggers a bounded revision step.

3.2 Framer Agent: Stakeholder-Aware Planning

The *Framer* is the first role-specific LLM module executed within each stakeholder pathway. It receives the shared *ExplanationCard* and transforms its evidence into a structured narrative plan, named *FramingTrace*. This artifact defines how the local explanation should be communicated before natural-language generation, including the organization of the narrative, the feature presentation order, the intended tone, the required cautions, and the operational instructions for the downstream *Narrator*.

Stakeholder adaptation is mainly introduced at this planning stage. For clinicians, the *Framer* can preserve a more medically detailed explanation; for patients, it prioritizes readability, uncertainty framing, and non-alarming formulations; for data scientists, it can retain more explicit references to attribution direction, feature ranking, relative magnitude, and model behavior. The *FramingTrace* remains an internal planning artifact rather than a user-facing explanation, thereby separating explanation planning from narrative realization.

3.3 Narrator Agent: Controlled Narrative Generation

The *Narrator* is the role-specific LLM module responsible for generating the explanatory text. It receives the original *ExplanationCard*, providing the factual and attributional evidence, alongside the *FramingTrace*, which defines the stakeholder-oriented narrative plan. The output is a candidate narrative grounded in the card and constrained by the *Framer*’s specified communicative strategy.

The generation process is controlled by the interaction between these two artifacts. The *ExplanationCard* constrains *what* can be said, including the prediction, probability, selected features, observed values, attribution values, and contribution directions. The *FramingTrace* constrains *how* this evidence should be communicated, including order, level of detail, tone, terminology, and cautions. This division keeps the narrative anchored to the original evidence while adapting its presentation to the target stakeholder.

The *Narrator* must preserve attribution meaning and respect the feature-level interpretations and generation constraints encoded in the *ExplanationCard*, including restrictions on causal, diagnostic, prognostic, and treatment-oriented claims. Its output remains provisional until it is evaluated by the *Reviewer*, preserving a separation between narrative generation and validation.

3.4 Reviewer Agent and Traceability

The *Reviewer* acts as the internal quality-control module for each stakeholder-specific pathway, receiving the *ExplanationCard*, the *FramingTrace*, and the candidate narrative before final acceptance.

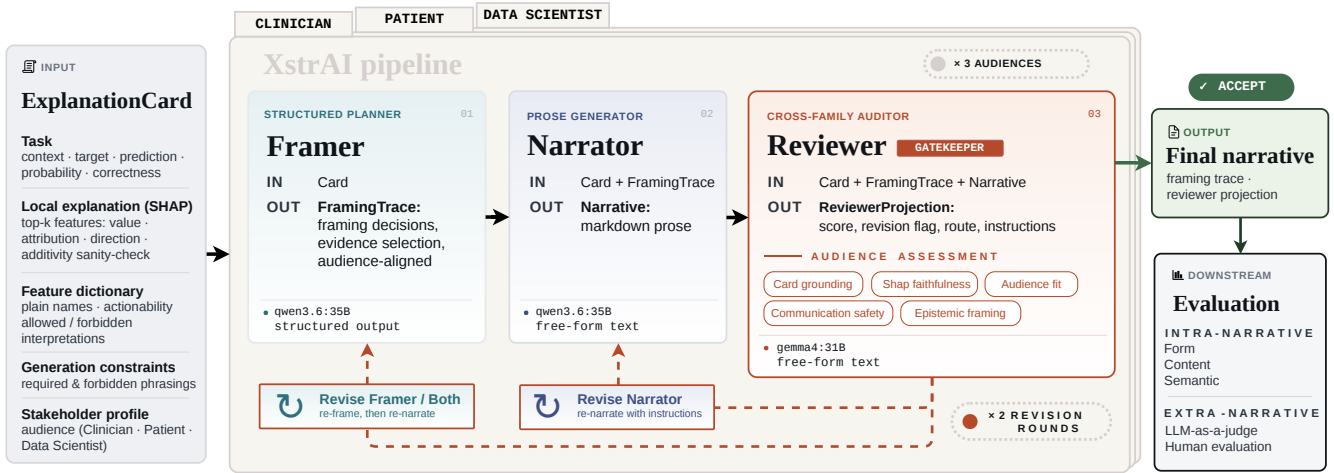


Figure 1: XstrAI architecture. Each case is represented as a structured ExplanationCard and processed through three stakeholder-specific pathways, each comprising three agents: the *Framer* produces a narrative plan, the *Narrator* generates the explanation, and the *Reviewer* evaluates the output across five audience-aligned dimensions. Rejected outputs trigger a bounded revision loop routed back to the Framer or Narrator. Finally, the generated narratives undergo intra- and extra-narrative evaluation.

Its assessment covers five shared dimensions across stakeholder-specific reviewers (top box in Figure 2): card grounding, SHAP faithfulness, epistemic framing, communication safety, and audience fit. Together, these checks are designed to flag potential deviations from the input evidence, attribution meaning, epistemic boundaries, communication-safety constraints, and stakeholder-appropriate register. For each candidate narrative, the Reviewer produces a structured assessment (bottom box in Figure 2) that includes an acceptance judgment, a score, a rationale, optional revision instructions, and, when needed, routing recommendations for refinement. Accepted narratives are returned as final outputs, while rejected ones trigger a bounded refinement loop involving the Framer, the Narrator, or both, depending on the identified issues. Updated artifacts are subsequently re-evaluated by the Reviewer, and the process iterates until acceptance or until the maximum number of revision rounds is reached. In the current implementation, each pathway can undergo at most two refinement iterations, yielding up to three generated narratives for the same case. This bound is consistent with empirical observations on multi-agent LLM systems, where consensus across iterations becomes substantially stable within a few rounds [8].

Distinct from the external evaluators used in the experimental assessment, the Reviewer serves as an internal quality-control component during inference. To support traceability, the framework logs the intermediate and final artifacts generated during execution, including inputs, narratives, review outputs, revision decisions, and final accepted explanations. This trace makes it possible to reconstruct the generation process, inspect failure modes, compare stakeholder pathways, and audit the system beyond the final text.

4 Experimental Setting

We conducted extensive experiments to evaluate our proposed approach. This section details the implementation of XstrAI, the selected baselines, and the evaluation protocol. All experiments

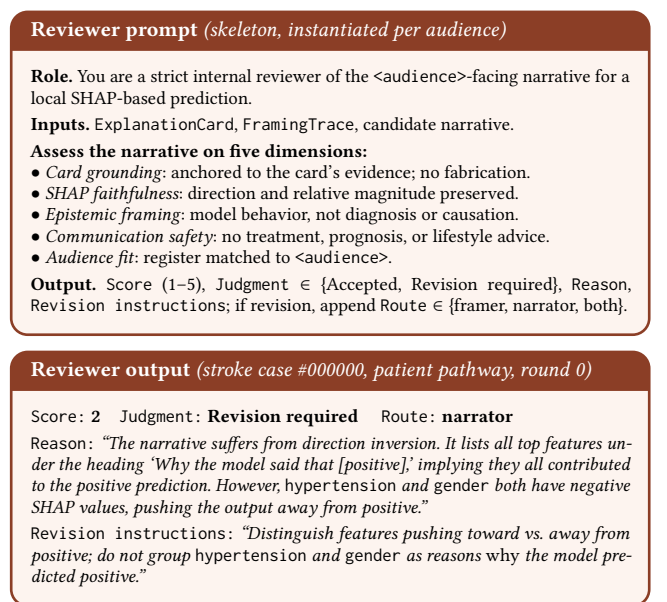


Figure 2: The Reviewer in action. *Top:* skeleton of the audience-conditioned Reviewer prompt; the same five dimensions are instantiated for Clinician, Patient, and Data Scientist. *Bottom:* a real Reviewer output that catches a SHAP direction inversion in a patient narrative and routes the revision back to the Narrator, preserving the existing FramingTrace.

were implemented in Python and executed on an Ubuntu 22.04 workstation equipped with an NVIDIA H100 GPU. The framework is model- and backend-agnostic: agent roles are defined at the prompt and schema level and can be served by any LLM endpoint

exposing chat completions, including remote APIs. To support reproducibility and external inspection, the repository includes the full framework implementation, stakeholder-specific prompts, reviewer configurations, evaluation pipelines, and experimental settings: <https://anonymous.4open.science/r/XstrAI-5528>.

4.1 Dataset, Predictive Models and Explanations

Our experiments involved two publicly available classification datasets: the Diabetes Dataset¹ and the Stroke Prediction Dataset². These datasets were selected because they offer interpretable tabular features, and a manageable feature space for assessing how local explanations can be transformed into stakeholder-specific narratives, consistently with recent work on XAI narrative generation [18].

The Diabetes Dataset comprises 768 records from the National Institute of Diabetes and Digestive and Kidney Diseases, with eight features and a binary diagnosis. The Stroke Prediction Dataset contains 5110 records with 10 features predicting stroke risk; preprocessing included one-hot encoding for categorical features (*gender*, *ever_married*, *Residence_type*), median imputation of missing BMI values, and discarding the *smoking_status* feature due to excessive missing data. For both datasets we addressed class imbalance by randomly undersampling the majority class (yielding 536 and 498 samples respectively), applied an 80%/20% train-test split, and normalized all features to the [0, 1] range.

For the classification task, we trained a Random Forest (RF) model on each dataset with default hyperparameters, achieving an accuracy of 0.68 (diabetes) and 0.77 (stroke). To interpret RF predictions, we computed local SHAP attributions [32], selected for their theoretical guarantees [13, 32] and its extensive validation in the medical domain [41]. We used the *TreeExplainer* algorithm from the Python SHAP library³, which provides efficient computation for tree-based models.

4.2 Baselines

We compare XstrAI against five baselines: Explingo [51], the most recent loop-based narrator for tabular SHAP attributions, and four controlled variants B1–B4 of XstrAI that incrementally introduce the components of our pipeline, given the limited availability of directly comparable systems for audience-aware narrative generation from fixed local attribution evidence.

Explingo. Explingo [51] introduces a dual-agent framework consisting of a narrator and a reviewer, which the original authors refer to as a grader. First, the narrator converts an explanation tuple into a short text using three labeled few-shot exemplars. Subsequently, the reviewer evaluates this generated output across four specific rubrics: accuracy, completeness, fluency, and conciseness. If the resulting average score falls below a predefined threshold, the reviewer automatically triggers a rewrite process. To enable a fair evaluation, we re-implement it under the same setup adopted for XstrAI, pairing a qwen narrator with a gemma reviewer, instead of using OpenAI APIs. Prompt assembly, rubric wording and the deterministic conciseness formula follow the original release. The

framework is audience-agnostic at every level: it returns a single narrative per instance, the narrator-reviewer exchange is limited to a scalar score and short free-text feedback, and the rubrics do not encode any reader-specific criterion. These properties motivate the controlled baselines that follow.

Incremental baselines B1–B4. We implement a family of incremental baselines that ablate the individual contributions of XstrAI, with consecutive baselines differing by a single design decision so as to isolate its effect. **B1** is the minimal configuration: a single language model receives the explanation card and produces one generic narrative, with neither audience conditioning nor any revision step. **B2** introduces the simplest form of audience awareness, a role description embedded in the prompt, while preserving the single-shot character of B1. **B3** adds a revision loop on top of B1: a narrator generates, a separate reviewer emits a 1-5 score together with an accept/revise verdict and a short comment, and the narrator rewrites until acceptance or until a fixed budget is exhausted. **B4** extends B3 with a rule-based verifier that examines machine-readable properties of the candidate narrative, section structure, presence of the numeric SHAP values in the prose, absence of list-like or JSON-like formatting, and a rewrite is requested unless both the reviewer and the verifier accept the output.

Implementation details. Every system consumes the same predictive model output and the same SHAP attribution, i.e. the same ExplanationCard, and every language model is instantiated from either Qwen3.6:35B [11] or Gemma4:31B [15]. For single-agent baselines (B1 and B2), we instantiate two versions of each system, one for each type of backbone, so that the comparison does not depend on a single language model. For configurations involving a Reviewer Agent (XstrAI, Explingo, B3, and B4), we use Qwen3.6:35B as the generator and Gemma4:31B as the reviewer, so as to avoid potential family bias during the review process [37, 46, 50]. When accounting for backbone and audience variants, the five baseline families expand into 11 distinct configurations, which together with the three audience-specific XstrAI variants constitute the 14-system pool used in the evaluation.

We denote systems as `system-backbone:audience`, where backbones are qwen or gemma and audience codes are C (Clinician), P (Patient), or DS (Data Scientist). Audience-agnostic or single-backbone systems omit the respective segments. For instance, B1-qwen denotes B1 instantiated on qwen, while B2-gemma:P denotes B2 instantiated on gemma for the Patient audience.

4.3 Evaluation framework

Audience-aware XAI narratives are not yet supported by a dedicated evaluation protocol covering both their observable textual properties and how they are received by external evaluators. We therefore designed a dedicated framework articulated along two regimes: an Intra-narrative module, computing observable properties of the narratives themselves, and an Extra-narrative module, eliciting judgments from external evaluators.

4.3.1 Intra-narrative Evaluation Module. This module comprises a diverse set of metrics to evaluate the narratives along three distinct axes: form, content, and semantic.

Form. To quantitatively assess the stylistic properties of the generated narratives, we performed a form-oriented analysis focusing

¹<https://www.kaggle.com/datasets/mathchi/diabetes-data-set>

²<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

³<https://shap.readthedocs.io/en/latest/>

on readability, lexical richness, verbosity, and repetition patterns. The objective of this analysis is not to evaluate factual correctness, but rather to characterize the structural and stylistic properties of the generated narratives, analyzing how different systems organize explanations and whether these characteristics change when narratives are adapted to specific stakeholders.

Specifically, we compute four complementary metrics. The average number of generated words per narrative (*#Words*) is used to estimate verbosity and explanation granularity, indicating how concise or detailed the generated explanations are. Readability is evaluated through the *Flesch–Kincaid Grade Level (FKGL)* [12], which estimates the educational level required to understand a text, with lower values indicating simpler language and higher values greater linguistic complexity. To evaluate lexical richness, we compute the *Moving-Average Type–Token Ratio (MATTR)* [5], a lexical diversity metric robust to text-length variation, where higher values indicate richer language use. Finally, repetitive generation patterns are quantified through the *Repeated Trigram Ratio (Rep-3)* [14], defined as the proportion of repeated three-word sequences appearing within a narrative. Higher values indicate stronger reliance on repetitive or templatic phrasing, while lower values suggest more diverse and fluent text generation.

All metrics are computed independently for each generated narrative and subsequently averaged across both datasets. Since some systems explicitly support stakeholder-oriented generation while others produce only generic explanations, the analysis additionally investigates whether stylistic properties vary when narratives are adapted to clinician, data scientist, or patient audiences.

Content. To evaluate the content faithfulness of the generated narratives, we compare the information extracted from each narrative against the corresponding ExplanationCard, which provides the reference feature values, SHAP contributions, and feature ordering.

We compute three agreement metrics introduced by Krishna et al. [26]. *Value Agreement (VA)* measures whether the feature values mentioned in the narrative match those in the ExplanationCard. *Sign Agreement (SA)* measures whether the direction of each feature contribution is preserved, i.e., whether a feature is described as raising or lowering the positive class consistently with the sign of its SHAP value. *Rank Agreement (RA)* measures whether the ordering of features in the narrative is consistent with the original feature ordering in the ExplanationCard.

To compute these metrics, we use two complementary extraction strategies. The first is a deterministic rule-based extractor that identifies feature mentions through canonical names, normalized variants, and predefined aliases. For each detected feature, the extractor analyzes a local text window to identify feature values, SHAP values, contribution direction, and rank indicators. Directions are inferred either from explicit SHAP values or from lexical markers such as *raises*, *lowers*, *positive contribution*, and *negative contribution*. Feature and SHAP values are extracted through numeric regular expressions and matched against the expected values in the ExplanationCard. Ranks are inferred from explicit ordinal markers such as *first*, *second*, or *most important*; otherwise, ranks are assigned according to feature appearance order.

The second strategy uses an LLM-based extractor based on GPT-5.4-mini, which infers feature mentions, contribution directions, and feature ordering from the narrative text. Compared to

deterministic extraction, this approach is more robust when attribution information is conveyed implicitly rather than through explicit SHAP terminology or rank markers.

The two extraction strategies are intentionally complementary: deterministic extraction provides a fully reproducible and easily verifiable pipeline, while the LLM-based extractor provides stronger semantic interpretation capabilities.

Finally, each feature is annotated as either *actionable* or *non-actionable*. Using these annotations, we compute the average rank displacement between the original feature ranking and the generated narrative. Positive displacement indicates feature promotion, while negative displacement indicates demotion. This allows us to assess whether stakeholder-oriented narratives prioritize actionable information over strict adherence to the original SHAP ranking.

Semantics. To quantify the semantic alignment of a generated narrative with a specific target audience, we developed a reference-based evaluation approach inspired by BERTScore [49]. We first compiled a corpus of representative texts tailored to our three user profiles: clinicians, data scientists, and patients. For clinicians and data scientists, we extracted the "Results" and "Discussion" sections from scientific papers published between 2024 and 2025 on PubMed and IEEE Xplore, respectively. The search queried titles containing "diabetes" or "stroke" alongside "Machine Learning". For the patient profile, we sourced informative lay-language medical articles from MedlinePlus⁴, the public health information platform maintained by the U.S. National Institutes of Health (NIH). After stripping markdown artifacts and citation patterns from the collected texts, we split the narratives into individual sentences and computed their embeddings using Sentence-BERT [40]. This yielded a domain-specific reference database comprising three stakeholder-specific embedding sets for each disease.

To ensure the quality of the reference data, we refined each set by removing noisy observations. Specifically, we computed a local density score for each sentence embedding based on the mean cosine similarity to its $k = 10$ nearest neighbors. Embeddings with a density Z-score outside the $[-2, 2]$ range were discarded as outliers. Finally, to prevent bias during evaluation, we randomly under-sampled the filtered sets so that all three stakeholder databases within a given disease domain contained the exact same number of embeddings.

To evaluate a generated narrative, we embed all its constituent sentences and project them into the corresponding disease’s reference space. For each sentence embedding, we retrieve the $k = 10$ nearest neighbors across the combined stakeholder databases. We then calculate the proportion of neighbors belonging to each audience class, averaged over all sentences in the narrative. The narrative is classified into the stakeholder category with the highest average score. If this assigned category matches the intended target audience, the prediction is considered correct. We refer to this metric as the Semantic Accuracy and applied the evaluation exclusively to XstrAI and the B2 baselines, as they target multiple stakeholders.

4.3.2 Extra-narrative Evaluation Module. While the Intra-narrative Evaluation Module assesses the intrinsic properties of the generated texts, the Extra-narrative Evaluation Module investigates the practical utility and subjective reception of the narratives from the

⁴<https://medlineplus.gov/diabetes.html>

perspective of the user. To achieve a comprehensive and robust assessment, we structured this module around a dual evaluation strategy: an LLM-as-a-Judge approach and a Human Evaluation through surveys.

LLM-as-a-judge. In the LLM-as-a-judge paradigm [4, 29, 50], a Large Language Model is employed to assess system outputs. In our study, the judge evaluates the narratives generated by XstrAI and the baselines through two complementary approaches, which we refer to as Re-identification Analysis and Competitive Ranking Analysis. For these evaluations, we leveraged three distinct generative models: Claude Sonnet 4.6, GPT-5.5, and Gemini 3.1. The judge models do not share the same architectural family as the qwen or gemma backbones, structurally avoiding self-preference bias [37, 46, 50].

Re-identification Analysis: To validate audience targeting, the three narratives must differ substantively in content and framing, not merely in surface markers such as headings or tone. In this task, the judge receives XstrAI’s three narratives in anonymised slots and assigns each to its intended audience. Before evaluation, narratives are sanitized to remove any system-identifying markers while preserving all prose, style, quantitative values, and SHAP terminology. The three narratives are shuffled randomly within each case, so a fixed-position guesser achieves chance baseline. Judges cannot abstain or mark ties. Across 3 judges and 208 cases, each judge evaluates 624 narrative trials.

Competitive Ranking Analysis: For a given sample and target audience, each judge ranks all 14 narratives produced by the evaluated systems from best to worst. This task is performed independently for each of the three target audiences and across all samples in both datasets, yielding a total of 1,872 ranking tasks and 26,208 rank observations.

We aggregate rankings into an Elo-style score using the Bradley-Terry model [3, 21], which expresses system quality as a continuous score based on pairwise win probability: a higher-scored system wins head-to-head comparisons more often. This approach is standard in LLM leaderboards [50]. Since our pool has no canonical reference, we centre scores at the pool mean, so 0 represents the average system, and positive/negative values denote above-and below-average performance. To determine whether differences among systems are statistically significant, we apply the Friedman omnibus test to assess whether ranking differences exist overall, and the Nemenyi post-hoc pairwise comparison [7] to identify which specific pairs differ significantly.

Finally, to evaluate the agreement among different judges, we compute the pairwise Kendall’s τ [23] and the ordinal Krippendorff’s α [25]. The latter is a multi-rater coefficient that aggregates the consensus of all three judges into a single metric per target⁵. For both metrics, higher values indicate a stronger consensus among the LLM judges.

Human Evaluation. To assess how human audiences perceive the XstrAI narratives, we conducted a preliminary evaluation through structured surveys. We randomly selected 3 instances from each dataset, resulting in a set of 6 tailored narratives per stakeholder profile (3 for Stroke, 3 for Diabetes). We recruited 15 human evaluators, divided evenly across the three audience groups (5 participants

Table 1: Evaluation dimensions and their corresponding survey statements for the patient audience. Participants rated each statement on a 5-point Likert scale.

Dimension	Statement
Understandability	I understood the factors influencing the model prediction.
Clarity	The explanation was clear and easy to read.
Usefulness	This explanation would help me better understand my health situation.
Trust	The explanation increased my trust in the AI system.
Cognitive Load	The explanation required too much effort to understand.

Table 2: Form-analysis metrics averaged across both datasets. Lower/higher values are preferred as indicated by the arrows.

Model	Audience	#Words	FKGL ↓	MATTR ↑	Rep-3 ↓
B1-gemma	Generic	166.56	10.76	0.628	0.206
B1-qwen	Generic	214.68	14.84	0.624	0.238
B2-gemma	Clinician	157.45	11.42	0.704	0.117
B2-gemma	Data Scientist	177.84	8.54	0.640	0.189
B2-gemma	Patient	153.93	10.35	0.782	0.081
B2-qwen	Clinician	191.01	14.88	0.716	0.135
B2-qwen	Data Scientist	269.44	14.07	0.668	0.163
B2-qwen	Patient	162.13	12.79	0.812	0.122
B3	Generic	215.63	13.66	0.601	0.362
B4	Generic	172.58	10.98	0.574	0.296
Explingo	Generic	38.38	15.37	0.792	0.001
XstrAI	Clinician	150.77	17.06	0.787	0.129
XstrAI	Data Scientist	232.23	15.69	0.752	0.042
XstrAI	Patient	161.63	12.09	0.806	0.064

per group), to evaluate the narratives corresponding to their specific profile. The clinician group consisted of medical specialists with at least three years of clinical experience, the data scientist group of professionals with at least three years of experience in data analysis and machine learning, and the patient group of lay participants without medical or technical backgrounds. The assessment covered five dimensions: Understandability, Clarity, Usefulness, Trust, and Cognitive Load. The Cognitive Load dimension was adapted from the NASA Task Load Index (NASA-TLX) [17], while the remaining items were derived from the Explanation Goodness Checklist and Explanation Satisfaction Scale [19]. All dimensions were slightly adapted to fit our scenario and rated on a 5-point Likert scale (ranging from 1 = Strongly Disagree to 5 = Strongly Agree). Table 1 details the specific dimensions and statements presented to the patient group. Finally, to mitigate potential ordering bias, both the sequence of the medical domains (Diabetes and Stroke) and the presentation order of the narratives within each domain were fully randomized for each participant.

5 Results and Discussion

5.1 Intra-narrative analysis

⁵Gwet’s AC2 with quadratic ordinal weights [16] aligns with α within ± 0.001 .

5.1.1 Form Analysis. Table 2 reports the stylistic and readability-oriented metrics computed across the generated narratives. Overall, the analysis reveals substantial differences in verbosity, readability, lexical diversity, and repetition patterns across both systems and stakeholder configurations.

The results show that audience adaptation affects both the content and style of the generated explanations. Patient-oriented narratives generally exhibit lower linguistic complexity than clinician- and data scientist-oriented ones. For example, XstrAI:P achieves a lower FKGL score (12.09) than XstrAI:DS (15.69), while data scientist narratives tend to be longer and more detailed.

Narrative repetition also varies substantially across systems. B3 and B4 exhibit the highest repeated trigram ratios (0.362 and 0.296, respectively), together with comparatively low MATTR scores, suggesting stronger reliance on template-based generation patterns. In contrast, XstrAI maintains consistently low repetition across all stakeholders, particularly for the data scientist target (0.042), despite generating longer narratives. This suggests that iterative refinement may help reduce boilerplate generation.

Lexical diversity further supports this interpretation. Our approach achieves consistently high MATTR scores across all targets, particularly for patient-oriented narratives (0.806), suggesting a more varied language use. Explingo also achieves high MATTR values with almost zero repetition; however, this behavior is likely influenced by its extremely short outputs (~ 40 words on average) rather than by a richer discourse structure. Indeed, despite its brevity, Explingo exhibits one of the highest FKGL scores in the benchmark.

The comparison between B2 and XstrAI also provides insights into the role of multi-stage refinement. B2 already demonstrates that simple role prompting is sufficient to induce measurable stylistic adaptation across stakeholders. However, XstrAI generally achieves higher lexical diversity together with lower repetition, suggesting that iterative refinement may help generate more natural and less templatic narratives.

Overall, XstrAI combines stakeholder adaptation, lexical richness, and low repetition while maintaining substantially longer narratives than compressed approaches like Explingo.

5.1.2 Content Analysis. Table 3 reports the Value Agreement (VA), Sign Agreement (SA), and Rank Agreement (RA) scores computed using both deterministic and LLM-based extraction approaches.

Regarding Value Agreement, most systems achieve nearly perfect scores under both extraction strategies, indicating that feature values are generally preserved correctly whenever they are explicitly mentioned in the narrative. The main exception concerns patient-oriented explanations, where raw feature values are intentionally suppressed. Interestingly, the B2 patient configuration still achieves non-negligible VA scores (0.650/0.782), suggesting that role prompting alone only partially suppresses explicit feature values. In contrast, XstrAI:P consistently suppresses explicit value mentions, indicating that the refinement stages enforce stakeholder-oriented communication constraints more effectively.

For both Sign Agreement and Rank Agreement, the LLM-based extraction strategy generally achieves higher scores than deterministic extraction, particularly for Explingo and patient-oriented narratives. In particular, SA scores obtained with LLM extraction

Table 3: Content-analysis metrics averaged across both datasets. Scores are reported as Deterministic (D) / LLM (L).

Model	Audience	VA (D/L)	SA (D/L)	RA (D/L)
B1-gemma	Generic	0.995 / 0.995	0.972 / 0.995	0.996 / 0.997
B1-qwen	Generic	1.000 / 1.000	0.972 / 1.000	0.979 / 1.000
B2-gemma	Clinician	1.000 / 1.000	0.917 / 1.000	0.943 / 0.971
B2-gemma	Data Scientist	1.000 / 1.000	0.985 / 1.000	0.985 / 0.988
B2-gemma	Patient	0.972 / 0.972	0.563 / 0.991	0.888 / 0.891
B2-qwen	Clinician	1.000 / 1.000	0.947 / 1.000	0.963 / 1.000
B2-qwen	Data Scientist	1.000 / 1.000	0.989 / 1.000	0.991 / 1.000
B2-qwen	Patient	0.650 / 0.782	0.337 / 0.992	0.926 / 0.961
B3	Generic	1.000 / 1.000	0.735 / 1.000	1.000 / 1.000
B4	Generic	1.000 / 0.999	0.802 / 1.000	0.993 / 0.999
Explingo	Generic	1.000 / 0.994	0.075 / 0.999	0.855 / 0.871
XstrAI	Clinician	1.000 / 0.998	0.979 / 1.000	0.897 / 0.989
XstrAI	Data Scientist	0.995 / 0.994	0.892 / 1.000	0.953 / 0.988
XstrAI	Patient	0.000 / 0.000	0.376 / 0.698	0.773 / 0.789

Table 4: Average feature rank displacement from the original ranking. Positive values indicate promotion in the generated narrative, negative values demotion.

Model	Audience	Actionable	Non-actionable
XstrAI	Clinician	0.011	-0.006
XstrAI	Data Scientist	0.004	0.013
XstrAI	Patient	0.622	-0.305

are close to or equal to 1.0 in most configurations, indicating that the direction of feature contributions is usually preserved even when attribution information is conveyed implicitly. Deterministic extraction, in contrast, produces lower scores in settings where explanatory information is expressed through free-form language rather than explicit SHAP terminology or numeric values.

Rank Agreement also reveals the impact of narrative rewriting on explanation structure. Explingo exhibits the lowest RA values (0.855/0.871), suggesting that aggressive rewriting substantially alters the ordering of explanatory evidence. XstrAI instead maintains stronger alignment with the original SHAP ranking for clinician and data scientist narratives, while patient-oriented explanations intentionally promote *actionable* features, i.e., features that the patient can potentially modify or intervene on. For example, body mass index is considered actionable, whereas age is not. This design prioritizes intervention-oriented communication over strict adherence to the original SHAP ordering.

To validate this hypothesis, Table 4 reports the average displacement of actionable and non-actionable features with respect to their original ranking. A clear divergence emerges for patient-oriented narratives: actionable features are promoted on average by +0.622 positions, while non-actionable features are demoted by -0.305. This effect is nearly absent for clinician and data scientist narratives, confirming that the patient-oriented strategy systematically prioritizes actionable information.

5.1.3 Semantic Analysis. The results of the semantic evaluation are detailed in Table 5. Across both datasets, XstrAI significantly

Table 5: Semantic accuracy across datasets and stakeholder targets. The best results for each dataset are in bold.

Dataset	Approach	Patient	Clinician	Data Scientist	Overall
Stroke	B2-gemma	0.04	0.29	1.00	0.44
	B2-qwen	0.01	0.24	1.00	0.42
	XstrAI	0.79	0.95	0.92	0.89
Diabetes	B2-gemma	0.27	0.46	0.50	0.41
	B2-qwen	0.21	0.57	0.27	0.35
	XstrAI	0.99	0.94	0.56	0.83

outperforms the B2 implementations (gemma and qwen) for the Patient and Clinician audiences. Notably, the baseline models struggle severely to adapt their language for the Patient, yielding semantic accuracies consistently below 0.30 on the diabetes dataset and below 0.10 on the stroke dataset.

B2 achieves perfect accuracy (1.00) on the stroke dataset for the Data Scientist, slightly surpassing XstrAI (0.92). However, the full approach maintains a much more robust and balanced performance across all stakeholder profiles. Consequently, the overall accuracy of XstrAI is more than double that of the baselines on both datasets (0.89 vs. 0.44/0.42 for stroke, and 0.83 vs. 0.41/0.35 for diabetes).

These findings highlight a critical limitation of standard audience-aware generation: simply embedding a role description into an LLM prompt (as done in B2) is insufficient to produce narratives with accurately tailored semantics. This gap is especially pronounced for the patient stakeholder, who represents an audience external to the professional medical and technical domains. Conversely, the introduction of a Reviewer Agent in our architecture appears effective, enabling both the generation and the targeted refinement of narratives which semantically match their intended stakeholders.

5.2 Extra-Narrative Analysis

5.2.1 LLM-as-a-judge: Re-identification Analysis. All three judges consistently assigned every generated narrative to its intended audience on both datasets, reaching 100% per-narrative accuracy and 100% triplet full-match. These scores are well above the corresponding random baselines of 33% and 16.7%. This suggests that XstrAI does not merely alter surface markers such as headings or tone, but produces narratives whose content and framing are consistently recognizable as audience-specific.

While these scores show that our approach tailors narratives to their intended audiences, they may also reflect the advanced reasoning capabilities of the state-of-the-art LLMs acting as judges, combined with the relative simplicity of the discriminative task. To provide a more rigorous and granular evaluation, we complement this analysis with the Competitive Ranking Analysis, directly comparing the quality of XstrAI against the baselines in a competitive scenario.

5.2.2 LLM-as-a-judge: Competitive Ranking Analysis. This analysis evaluates whether systems are preferred by LLM judges when narratives are ranked for a specific target audience. Figure 3 reports Bradley-Terry Elo scores with confidence intervals per audience; Table 6 reports inter-judge agreement.

Table 6: Inter-judge agreement on the Competitive Ranking Analysis. Pairwise mean Kendall τ (averaged over the two datasets) and ordinal Krippendorff’s α per dataset (multi-rater across all 3 judges). Higher = more agreement; the shaded row falls below the 0.80 reliability cut-off.

Audience	Mean Kendall τ			Krippendorff α_{ord}	
	Claude GPT	Claude Gemini	GPT Gemini	diabetes	stroke
Patient	0.80	0.70	0.71	0.87	0.82
Data Scientist	0.85	0.72	0.73	0.90	0.86
Clinician	0.23	0.31	0.33	0.34	0.45

The Elo scores are associated with narrow error bars, indicating stable ranking differences. The Friedman test rejects the null hypothesis that all systems are equivalent ($p \ll 10^{-10}$ on every panel), and the Nemenyi pairwise comparison sets a critical difference of 0.79 rank positions, a threshold exceeded by most observed gaps.

XstrAI ranks first on two of its three matched audiences: *Clinician* (Elo +2.60, $\sim 85\%$ wins over B1-gemma) and *Patient* (Elo +6.27, $\sim 84\%$ wins over B2-qwen:P). On *Data Scientist*, it ranks third behind B2-qwen:DS (Elo +6.07) and B2-gemma:DS (Elo +4.56).

A consistent specialization effect is visible across all three panels. Audience-conditioned systems dominate their matched target but score poorly elsewhere: XstrAI:P, for instance, moves from +6.27 on *Patient* to -4.88 on *Data Scientist*. Generic baselines (B1, B3, B4) remain near the pool average regardless of target, confirming that audience-specific adaptation is a necessary component of narrative quality. Explingo ranks among the weakest overall, performing comparatively better on *Patient* alone, where its short and direct style partially satisfies accessibility requirements. Its consistently low scores on *Clinician* and *Data Scientist* show that concision alone cannot substitute for controlled, attribution-aware generation.

Audience-level interpretation. The three audiences reveal distinct patterns in how much pipeline complexity the task demands.

On *Clinician*, inter-judge agreement is the lowest of the three targets (Table 6), reflecting the legitimate diversity of acceptable clinical styles rather than evaluator noise. Yet it is precisely on this contested target that XstrAI:C achieves its widest margin in the benchmark: a +1.71 Elo lead over the runner-up B1-gemma, corresponding to a $\sim 85\%$ pairwise win probability and the largest first-to-second gap observed across all panels. The result suggests that multi-round refinement is most valuable where no single formulation dominates, allowing the pipeline to converge on outputs that single-prompt baselines, committed to one style, cannot match.

On *Patient*, XstrAI:P leads, followed by the baselines B2-qwen:P and B2-gemma:P. Explingo performs better here than on any other target, confirming that simplified, accessible language partially meets patient requirements. Even so, audience-conditioned systems rank above it, showing that simplicity alone is insufficient: a personalized and carefully framed explanation remains preferable to a merely concise one. The stronger inter-judge consensus on this target (ordinal $\alpha \geq 0.82$) confirms the stability of this ordering.

On *Data Scientist*, the dominant requirement is raw SHAP attribution density and quantitative completeness. A data scientist familiar with feature-attribution methods expects explicit values,

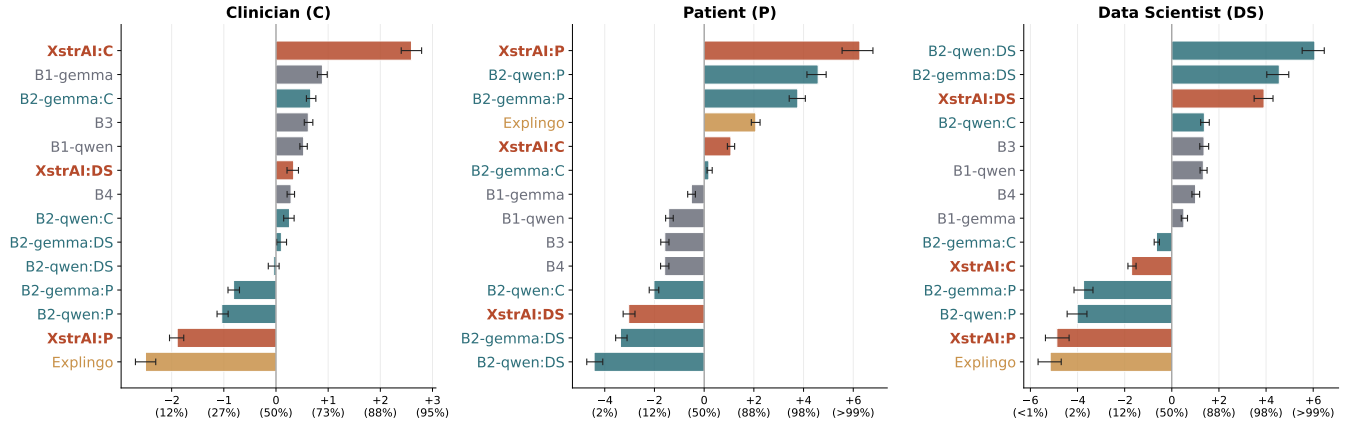


Figure 3: Competitive ranking analysis: Elo log-strength (zero-centered; 0 = average system) across three audiences, pooled over 3 judges and 2 datasets. Panels are sorted by Elo score, and percentages indicate pairwise win probability against the average system (e.g., Elo +1 $\approx$ 73%). Error bars represent 95% bootstrap confidence intervals.

Table 7: Average human evaluation scores on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Higher scores are better, except for Cognitive Load.

Dimension	Patient	Data Scientist	Clinician
Understandability	4.2	3.9	4.7
Clarity	3.8	3.8	4.3
Usefulness	3.5	3.6	4.3
Trust	3.0	3.3	4.1
Cognitive Load	2.2	2.4	2.4

rankings, and directions, properties that a single dense prompt naturally maximises. B2-qwen:DS and B2-gemma:DS lead accordingly, with XstrAI:DS close behind. All top-ranked systems are audience-conditioned, confirming that specialization remains necessary even when the narrative structure is simpler. The gap between audience-conditioned and generic systems is as large here as on any other target; what changes is the form that specialization should take.

5.2.3 Human Evaluation. Table 7 reports the average human evaluation scores across the five assessed dimensions for each target audience. Clinicians reported the highest scores among all stakeholders for Understandability (4.7), Clarity (4.3), Usefulness (4.3), and Trust (4.1), indicating that the generated narratives were exceptionally well-received and deemed highly supportive by medical professionals.

Interestingly, the Data Scientist group assigned more modest ratings across these dimensions, with average scores ranging between 3 (Neutral) and 4 (Agree). As detailed in Table 2, this group received the longest narratives, with an average of 232 words compared to 150 for clinicians. Although they processed longer texts, their reported Cognitive Load was identical to that of the medical audience (2.4). This suggests that the increased length did not translate into a significantly higher reading effort for technical users.

Finally, Patients found their tailored narratives to be highly understandable (4.2) and clear (3.8), while experiencing the lowest

Cognitive Load among all groups (2.2). However, there is room for improvement regarding Trust (3.0) and Usefulness (3.5), which received neutral-to-positive, yet limited, evaluations.

These findings suggest that XstrAI generates audience-specific narratives that are clear and require low cognitive effort. Given the limited number of participants, these findings should be interpreted as a preliminary but promising assessment of stakeholder-specific narrative perception rather than as a definitive usability validation.

6 Conclusion

This paper introduced XstrAI, a multi-agent framework for generating stakeholder-specific narratives from fixed local feature-attribution evidence. By separating explanatory evidence from its communication, XstrAI employs an immutable ExplanationCard and a Framer–Narrator–Reviewer pipeline to control how the same prediction explanation is selected, organized, verbalized, and assessed for different audiences.

Our evaluations show that XstrAI produces clearly differentiated narratives while showing strong empirical alignment with the underlying attribution evidence. Compared to generic and single-prompt baselines, the framework achieves lower repetition, higher lexical diversity, stronger audience fit, and competitive or superior performance in deterministic and LLM-based assessments. The findings also indicate that stakeholder adaptation may legitimately reshape explanation structure; for instance, in patient narratives, actionable features are prioritized over strict SHAP-rank ordering.

Future work will examine prompt-level ablations and the robustness of the framework under controlled noise in the input explanations. We also plan to assess whether the same communication layer design generalizes beyond SHAP to other XAI techniques and beyond tabular clinical prediction to other application contexts. Finally, larger and more targeted human evaluations will be needed to assess stakeholder-specific usefulness, trust, cognitive load, and communicative adequacy more deeply.

References

- [1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Túlio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *CHI*. ACM, 81:1–81:16.
- [2] Ahsan Bilal, David Ebert, and Beiyu Lin. 2025. Llms for explainable ai: A comprehensive survey. *arXiv preprint arXiv:2504.00125* (2025).
- [3] RALPH ALLAN BRADLEY and MILTON E. TERRY. 1952. RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS: THE METHOD OF PAIRED COMPARISONS. *Biometrika* 39, 3-4 (12 1952), 324–345. [arXiv:https://academic.oup.com/biomet/article-pdf/39/3-4/324/930466/39-3-4-324.pdf](https://academic.oup.com/biomet/article-pdf/39/3-4/324/930466/39-3-4-324.pdf) doi:10.1093/biomet/39.3-4.324
- [4] David Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *ACL (1)*. Association for Computational Linguistics, 15607–15631.
- [5] Michael A. Covington and Joe D. McFall. 2010. Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). *J. Quant. Linguistics* 17, 2 (2010), 94–100.
- [6] Pedro Henrique Luz de Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. 2025. Principled Personas: Defining and Measuring the Intended Effects of Persona Prompting on Task Performance. In *EMNLP*. Association for Computational Linguistics, 26857–26886.
- [7] Janez Demsar. 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* 7 (2006), 1–30.
- [8] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mor-datch. 2024. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In *ICML (Proceedings of Machine Learning Research)*. PMLR / OpenReview.net, 11733–11763.
- [9] Latifa Dwiyantri, Sergio Ryan Wibisono, and Hidetaka Nambo. 2025. Contextual-SHAP : Enhancing SHAP Explanations Through Contextual Language Generation. In *WSSE*. ACM, 231–237.
- [10] Upol Ehsan, Q. Vera Liao, Michael J. Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *CHI*. ACM, 82:1–82:19.
- [11] An Yang et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).
- [12] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.
- [13] Montgomery Flora, Corey Potvin, Amy McGovern, and Shawn Handler. 2022. Comparing explanation methods for traditional machine learning models part 1: an overview of current methods and quantifying their disagreement. *arXiv preprint arXiv:2211.08943* (2022).
- [14] Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A Theoretical Analysis of the Repetition Problem in Text Generation. In *AAAI*. AAAI Press, 12848–12856.
- [15] Google DeepMind. 2026. Gemma 4 Model Card. https://ai.google.dev/gemma/docs/core/model_card_4
- [16] Kilem Li Gwet. 2008. Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement. *Brit. J. Math. Statist. Psych.* 61, 1 (2008), 29–48. doi:10.1348/000711006X126600
- [17] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.
- [18] Yifan He and David Martens. 2026. An Agentic Approach to Generating XAI-Narratives. *arXiv preprint arXiv:2603.20003* (2026).
- [19] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257.
- [20] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiwu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *ICLR*. OpenReview.net.
- [21] David R. Hunter. 2003. MM algorithms for generalized Bradley-Terry models. *Annals of Statistics* 32 (2003), 384–406. <https://api.semanticscholar.org/CorpusID:17436309>
- [22] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38.
- [23] M. G. KENDALL. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika* 30, 1-2 (06 1938), 81–93. [arXiv:https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf](https://academic.oup.com/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf) doi:10.1093/biomet/30.1-2.81
- [24] Minjung Kim, Saebyeol Kim, Jinwoo Kim, Tae-Jin Song, and Yuyoung Kim. 2024. Do stakeholder needs differ? - Designing stakeholder-tailored Explainable Artificial Intelligence (XAI) interfaces. *Int. J. Hum. Comput. Stud.* 181 (2024), 103160.
- [25] klaus krippendorff. 2011. Computing Krippendorff’s Alpha-Reliability. (01 2011).
- [26] Satyapriya Krishna, Tessa Han, Alex Gu, Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju. 2024. The Disagreement Problem in Explainable Machine Learning: A Practitioner’s Perspective. *Trans. Mach. Learn. Res.* 2024 (2024).
- [27] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for “Mind” Exploration of Large Language Model Society. In *NeurIPS*.
- [28] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *CoRR* abs/2110.10790 (2021).
- [29] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *EMNLP*. Association for Computational Linguistics, 2511–2522.
- [30] Angela Lombardi, Francesca Arezzo, Eugenio Di Sciascio, Carmelo Arditto, Michele Mongelli, Nicola Di Lillo, Fabiana Divina Fascilla, Erica Silvestris, Anila Kardhashy, Carmela Putino, et al. 2023. A human-interpretable machine learning pipeline based on ultrasound to support leiomyosarcoma diagnosis. *Artificial Intelligence in Medicine* 146 (2023), 102697.
- [31] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lécué, Gianclaudio Malignieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Inf. Fusion* 106 (2024), 102301.
- [32] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *NIPS*. 4765–4774.
- [33] David Martens, James Hinns, Camille Dams, Mark Vergouwen, and Theodoros Evgeniou. 2025. Tell me a story! Narrative-driven XAI with Large Language Models. *Decis. Support Syst.* 191 (2025), 114402.
- [34] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267 (2019), 1–38.
- [35] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT*. ACM, 220–229.
- [36] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* 55, 13s (2023), 295:1–295:42.
- [37] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *NeurIPS*.
- [38] Joon Sung Park, Joseph C. O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *UIST*. ACM, 2:1–2:22.
- [39] Alun D. Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in Explainable AI. *CoRR* abs/1810.00184 (2018).
- [40] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410
- [41] Qiyang Sun, Alican Akman, and Björn W Schuller. 2025. Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare* 6, 2 (2025), 1–31.
- [42] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In *ACL (Findings)* (*Findings of ACL*). Association for Computational Linguistics, 599–621.
- [43] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *NeurIPS*.
- [44] Vittoria Vineis, Matteo Silvestri, Lorenzo Antonelli, Filippo Betello, and Gabriele Tolomei. 2026. PONTE: Personalized Orchestration for Natural Language Trust-worthy Explanations. *CoRR* abs/2603.06485 (2026).
- [45] Noah Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhao Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *ACL (Findings)* (*Findings of ACL*). Association for Computational Linguistics, 14743–14777.
- [46] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghui Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large Language Models are not Fair Evaluators. In *ACL (1)*. Association for Computational Linguistics, 9440–9450.
- [47] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents

- through Psychological Interviews. In *ACL (1)*. Association for Computational Linguistics, 1840–1873.
- [48] Matej Zecevic, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. *Trans. Mach. Learn. Res.* 2023 (2023).
 - [49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
 - [50] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS*.
 - [51] Alexandra Zyttek, Sara Pidò, Sarah Alnegheimish, Laure Berti-Équille, and Kalyan Veeramachaneni. 2024. Explingo: Explaining AI Predictions using Large Language Models. In *IEEE Big Data*. IEEE, 1197–1208.