Bridging the Domain Gap: Evaluating Fact-Grounded Knowledge Graph Narratives for Explainable AI in Clinical Decision Support

Valentin Göttisheim, Holger Ziekow,
Peter Schanbacher
Faculty of Business Information Systems
Furtwangen University
Furtwangen, Germany
e-mail: firstname.surename@hs-furtwangen.de

Djaffar Ould-Abdelsam
Université de Haute-Alsace
IRIMAS Laboratory, Université de Haute-Alsace
68100 Mulhouse, France
e-mail: djaffar.ould-Abdelsam@uha.fr

***Abstract*—Clinicians need transparent reasoning to trust Artificial Intelligence recommendations, but standard explanation methods lack clinical semantics. To address this, we transform an Onkopedia colon carcinoma guideline into a semantically enriched Knowledge Graph by segmenting text, extracting and merging semantic concepts, enriching gaps with registry data, and anchoring features to graph nodes. Using a predictive model, we compute Shapley Additive Explanations feature attributions and generate fact-grounded narratives via large language models that directly reference guideline evidence. We compare three contexts across 65 synthetic colorectal cancer cases (195 narratives) and find that KG-based narratives reduce hallucinations, speculation, and contradictions. Embedding KG-grounded narratives in clinical decision-support tools promises to shorten expert review cycles, surface guideline deviations, and bridge the explainability gap between data scientists and clinicians.**

Keywords-Keywords— Explainable Artificial Intelligence; XAI; Knowledge Graphs; Shapley Additive Explanations; SHAP; Narrative Generation; Claim Verification.

I. Introduction

Clinical decision support models promise early insights but often function as opaque black boxes [1]. Clinicians require transparent, evidence-based explanations to understand how input features drive predictions [2]. In practice, model development is a collaborative, iterative process: data scientists train and refine predictive models, generate interim explanations, and oncologists review these artifacts against clinical knowledge, suggest adjustments, and feed feedback into retraining until statistical performance and clinical relevance converge. This real-world feedback loop motivates our work.

To bridge the gap between raw model outputs and clinically meaningful interpretation, we augment Shapley Additive Explanations (SHAP) outputs with fact-grounded narratives linked to an authoritative guideline-derived Knowledge Graph (KG). Our contributions are threefold:

1) Extract and structure clinical guideline content into a semantically rich KG.
2) Compute SHAP attributions for model features and anchor them to KG nodes.
3) Generate narrative explanations referencing the KG, yielding traceable, domain-specific rationales.

Standard SHAP bar charts quantify feature influence but lack clinical semantics. By mapping attributions to KG nodes derived from colon carcinoma guidelines, our approach enriches explanations with medical context—enabling clinicians to reason in domain-specific terms and data scientists to identify discrepancies from accepted evidence. We therefore ask how such fact-grounded narratives affect four claim categories—*Hallucination*, *Contradiction*, *Speculation*, and *Extrapolation*:

(RQ1) Does KG anchoring reduce hallucinations?
(RQ2) Does KG anchoring reduce contradictions?
(RQ3) Does KG anchoring reduce speculative statements?
(RQ4) Does it keep extrapolations within the boundaries established by using guideline text alone?

If successful, this strategy could streamline expert review and facilitate the way for prospective clinical validation. The remainder of the paper is organized as follows: In Section III we present the proposed methods, including KG construction and narrative generation. Section IV reports quantitative and qualitative results. Section V discusses implications and limitations. Section VI concludes with future directions.

II. Related Work

Shapley values provide theoretically grounded, local feature attributions that have become standard in explainable clinical ML [3], but dense bar-chart displays impose high cognitive load on physicians [4]. To improve interpretability, template-based systems, such as *SHAPstories*, convert attributions into short rationales, yielding modest trust gains [5], while constrained decoding in *EXPLINGO* reduces hallucinations in general domains [6]. Burton et al. frame explanation verbalization as a data-to-text task with the TEXEN corpus—496 SHAP/LIME-to-narrative pairs—reporting factual error rates of 25%-42% for models like BART and T5 [7]. Although these methods enhance usability, they lack integration with domain-specific clinical knowledge.

Evaluation of explanation quality typically distinguishes between *faithfulness*—how accurately an explanation reflects the underlying model—and *plausibility*—how well it aligns with human judgment [8–10]. Kroeger et al. demonstrate that larger language models can yield less faithful post-hoc explanations without additional constraints [11], and Lanham

et al. offer a fine-grained benchmark for faithfulness in chain-of-thought reasoning [12]. Diagnostic probes, such as Walk-the-Talk and the FaithEval suite, complement traditional lexical overlap metrics (BLEU, ROUGE) by assessing deeper semantic and factual fidelity [13][14]. To build upon this strand, we introduce a structured factual-consistency framework that quantifies divergences across four categories: *Hallucination*, *Extrapolation*, *Speculation*, and *Contradiction*, as defined in Table II and applied in Table IV.

Knowledge Graphs enhance semantic structure, traceability, and bias control in otherwise opaque model explanations [15]. Typical KG construction pipelines involve text segmentation, entity and relation extraction, canonicalization, ontology alignment, and population [15], while widely used biomedical resources, like the UMLS Metathesaurus and Bio2RDF, integrate millions of curated concepts from diverse ontologies [16]. Domain-grounding systems, such as *XplainLLM*, anchor generated explanations in KG triples; *DR.KNOWS* integrates UMLS—a large compendium of biomedical terminologies—for diagnostic safety [17][18]. Cross-domain cybersecurity work highlights that LLM-based verbalization of SHAP tables can still wander off-fact without authoritative grounding [19]. Emerging LLM-based tools (e.g., Text2KG, LLM-Assisted Knowledge Graph Engineering) automate parts of these pipelines but face challenges, such as hallucination and schema drift [20][21]. Crucially, no existing approach constructs KGs directly from prescriptive clinical guidelines—a gap our guideline-driven pipeline addresses by extracting semantic concepts from Onkopedia guidelines, enriching them with registry data, and anchoring model features to KG nodes.

Building on post-hoc feature attributions (SHAP), narrative verbalization, domain-specific evaluation metrics, and established KG construction pipelines, we address the challenge of grounding model explanations in clinical evidence. We integrate guideline-derived Knowledge Graph construction with SHAP-anchored narrative generation to produce explanations that are both interpretable and verifiable. We evaluate factual accuracy by fact-checking statements in the generated narratives against patient case records and quantify divergences from the ground truth. This methodology yields fact-anchored narratives that clinicians can immediately verify against clinical guidelines, enhancing trust and accelerating prospective validation.

III. Proposed Methods

We developed an end-to-end pipeline that (i) transforms the Onkopedia colorectal-cancer (CRC) guideline [22] into a semantically enriched Knowledge Graph, (ii) computes Shapley Additive Explanations attributions on an XGBoost predictive model to quantify feature importance, and (iii) generates fact-grounded narrative explanations via large language models (LLMs), which we evaluate experimentally for factual consistency.

A. Knowledge Graph Representation

We represent the guideline-derived KG as a labeled directed graph, where nodes correspond to clinical semantic concepts (e.g., therapies, biomarkers, patient characteristics), edges denote typed relationships between them, and both nodes and edges carry labels derived from the medical guideline.

We implemented a six-stage pipeline to transform the CRC guideline into a semantically enriched Knowledge Graph:

Step 1: Preprocessing & Chunking: Clean raw guideline text (remove headers, footers) and segment into traceable 100-character chunks with metadata (chapter, page, hash).

Step 2: Concept & Relation Extraction: Apply GPT-o4-mini-high with structured prompts to extract semantic concepts as entities with attributes (name, description, confidence) and their inter-relations into a validated JSON schema.

Step 3: Subgraph Integration & Clustering: Merge chunk-level subgraphs into an initial graph, cluster entities by thematic category, consolidate identical identifiers, and link synonyms.

Step 4: Registry Enrichment: Identify missing clinical concepts, insert placeholder nodes, and enrich them with real-world CRC registry attributes (e.g., age, KRAS status, ECOG).

Step 5: Master Graph Assembly: Integrate all enriched subgraphs under a central root node, serialize in Markdown, and export to Neo4j format for queryability.

Step 6: Provenance Annotation: Attach detailed source metadata (document, chapter, page, chunk ID, hash) to every node and edge for auditability.

B. Narrative Generation

Based on a real-world colorectal-cancer registry data schema excerpt provided by our research partner, we built a simulation and generated 20,000 synthetic patient records. We trained an XGBoost model to forecast patient-level treatment decisions and quantified feature importance with SHAP contribution scores (ϕ_i) using the TreeExplainer algorithm [3]. SHAP decomposes each prediction $f(\mathbf{x})$ as:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{M} \phi_i,$$

where ϕ_0 is the model's expected output and each ϕ_i the marginal contribution of feature i. We linked features to their corresponding nodes in the guideline-derived KG, ensuring semantic grounding. However, not all features can be anchored to the KG, since some registry variables (e.g., body mass index or weeks since initial diagnosis) are not guideline-based clinical concepts. We then synthesized 65 colorectal-cancer patient personas—each defined by demographic variables, TNM stage, ECOG performance status, Charlson Comorbidity Index [23], and molecular biomarker profile—and stratified them into three complexity tiers: (i) uncomplicated cases without guideline conflicts; (ii) biomarker-driven cases; and (iii) multimorbid cases with conflicting recommendations. For each persona, we computed SHAP attributions using TreeExplainer on the

XGBoost predictive model and selected the ten highest-impact features by absolute SHAP magnitude. We then generated narrative explanations in three grounding contexts (OA, GL, KG), defined in Table I, with GPT-o4-mini-high, supplying both the complete patient CSV record and the top-ten SHAP features as patient case data. This 3 × 65 factorial design produced 195 narratives, enabling paired comparisons of factual consistency across grounding strategies. To evaluate the incremental impact of integrating clinical guidelines and Knowledge Graph information, we prompt the LLM (GPT-o4-mini-high) to generate narrative explanations under the three controlled contexts (OA, GL, KG). All narratives follow a standardized Markdown template to control for length and format, ensuring identical format and length constraints across experimental conditions.

TABLE I. GROUNDING CONTEXTS FOR NARRATIVE GENERATION

Context	Description
OA (Only-Attributes)	Patient case data alone, excluding guideline or KG context.
GL (Guideline)	Patient case data plus extracted guideline excerpts with explicit citations.
KG (Knowledge Graph)	Patient case data and full KG in Markdown, including labels, relations, and provenance.

C. Claim Extraction and Evidence Matching

We parsed each created narrative with GPT-o4-mini-high to extract individual asserted claims (complete sentences). For each claim, we matched its content against the patient case data (patient attributes and corresponding SHAP attributions). The LLM was prompted to flag each claim without direct support in the patient case data as *inferred* and to classify it into four categories: **Hallucination**, **Contradiction**, **Extrapolation**, and **Speculation**, as defined in Table II.

TABLE II. INFERRED CLAIM CATEGORIES AND DEFINITIONS

Category	Why the claim is inferred
Hallucination	The claim asserts a patient-specific fact that is *not present* in the case data or SHAP features; the model introduces new clinical information not observed in the input.
Contradiction	Claim **conflicts** with patient case data.
Extrapolation	Guideline-consistent generalization that lacks direct case evidence.
Speculation	Conjecture with insufficient grounding (not verifiable against case or guideline).

In the following, we illustrate examples of the LLM evaluated claim extraction and evidence matching phase. Each category in Table II is exemplified with excerpts from the LLM evaluation to illustrate the four distinct ways in which a generated *inferred* claim can arise. According to the **Extrapolation** criterion, a claim is clinically plausible and drawn from the guideline but lacks direct support in the patient record. For example:

> *"For a patient with stage I (T2 N0 M0) colon carcinoma, complete surgical resection is curative and no adjuvant chemotherapy is indicated."*

Here, the tumor stage (T2 N0 M0) is correctly taken from the case data, yet the recommendation about cure and omission of chemotherapy, while guideline-based, cannot be verified against any patient-specific attribute. Such extrapolations are nevertheless desirable, because they showcase the language model's ability to enrich its output with domain knowledge and provide broader narrative explanations rather than relying solely on SHAP-derived feature attributions. A **Speculation** covers plausible inferences that nonetheless lack explicit evidence. For example:

> *"ECOG 1 (–0.12) and a high comorbidity burden (CDRRHIGH_yes, –0.10) further lowered the probability because of toxicity concerns."*

Although ECOG and comorbidity are real features, attributing the SHAP-driven probability drop to "toxicity concerns" is conjectural and not encoded in the patient case. Such speculation are undesirable, as it introduces clinical reasoning not backed by case data and can mislead users about the true factors influencing the model. By contrast, a **Hallucination** arises when the model fabricates a patient-specific fact that does not appear in the input at all. Consider:

> *"Difference 1: According to the guidelines, an anti-EGFR antibody should be added for RAS-wild-type disease, whereas the model instead selects a BRAF-targeted agent (AB)."*

This statement wrongly attributes BRAF targeting to AB—a fact not mentioned in the case data. Such hallucinations are undesirable because they introduce clinical assertions not backed by case data, undermining trust in the explanation and potentially misleading downstream decisions.

Finally, **Contradiction** occurs when a claim directly conflicts with documented attributes. For instance:

> *"This 55-year-old man with resected rectal cancer (T3 N1 M1) and solitary liver and lung metastases has undergone complete surgical removal of all metastases."*

This contradicts the record's single-metastasis count (`NUMBER_METASTASES`=1) and notes R0 resection only for the primary tumor. Such contradictions are undesirable because they misrepresent case facts.

To validate claim extraction and evidence matching, which were performed automatically using the OpenAI GPT-o4-mini-high model, we randomly sampled 20 claims and computed classification accuracy with 95% Wilson-score confidence intervals to account for small-sample inference [24]. The LLM correctly classified 19 out of 20 cases (95% accuracy), yielding a Wilson 95% confidence interval of 76.4%–99.1%. Even at the lower bound, fewer than 25% of labels are expected to

be incorrect, justifying the use of automatic evaluation for the quantitative analyses.

IV. Results

To evaluate the factual consistency of the generated narratives across three grounding contexts—KG, GL, and OA—we report both quantitative counts and qualitative examples. Results are presented in three parts: overall observed vs. inferred claim counts, composition of inferred categories, and evaluation reliability.

A. Observed vs. Inferred Claims

We evaluated the generated narratives and labeled every asserted claim as either *observed* or *inferred*. A claim is *observed* when it is directly supported by the patient record (e.g., tumor stage or biomarker status) or explicitly grounded by a SHAP attribution that links a named feature to the model's prediction. A claim is *inferred* when it lacks such direct support; inferred claims were further categorized.

TABLE III. Overall observed vs. inferred claim counts by context

Context	Total	✓	❍	% Observed
KG	1 128	367	761	32.5 %
GL	1 125	243	882	21.6 %
OA	1 107	395	712	35.7 %

We report the proportion of observed versus inferred claims across the 195 narratives. Table III summarizes the total number of observed (✓) and inferred (❍) claims across the three grounding contexts. Narratives generated with KG grounding achieved 32.5 % observed claims (367/1 128), outperforming the GL context, which yielded only 21.6 % (243/1 125). The OA context performed comparably to KG with 35.7 % observed claims (395/1 107 vs. KG).

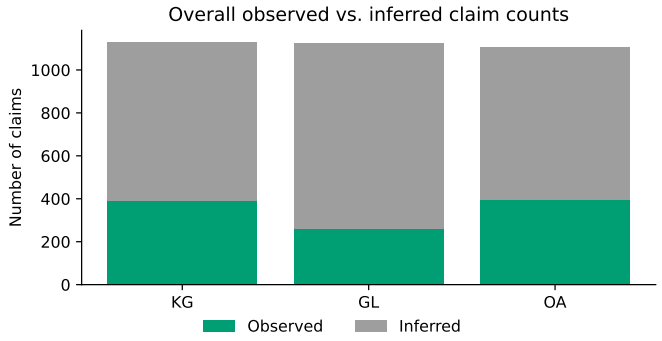

Figure 1. Overall observed vs. inferred claim counts by context (observed = case/SHAP-backed; inferred = not directly case-backed).

Figure 1 plots overall observed vs. inferred claim counts by context. Observed shares differed across the three contexts: explanations grounded in the KG achieved higher observed shares than those from the GL baseline, while OA and KG did not differ much. These findings indicate that KG-grounded input improves consistency over GL-context narratives, while OA may benefits from a narrower input scope with fewer opportunities for *inferred* claims.

B. Inferred Claim Categories

Table IV details the distribution of *inferred* claims by category—**Extrapolation**, **Speculation**, **Hallucination**, and **Contradiction**—expressed as a percentage of total claims in each context.

TABLE IV. Inferred claim category rates (percentage of total claims)

Category	KG	GL	OA
Extrapolation	64.8 %	73.7 %	61.9 %
Speculation	0.5 %	2.0 %	1.1 %
Hallucination	0.0 %	0.4 %	0.2 %
Contradiction	0.1 %	0.6 %	1.1 %

Extrapolation is the predominant inferred category across all contexts. However, the KG condition achieves substantial gains in factual precision and safety: no hallucinations were observed under this setup (0.0 %), speculation drops to 0.5 %, and contradictions fall to just 0.1 %. In contrast, the GL context shows higher rates of speculation (2.0 %) and contradiction (0.6 %).

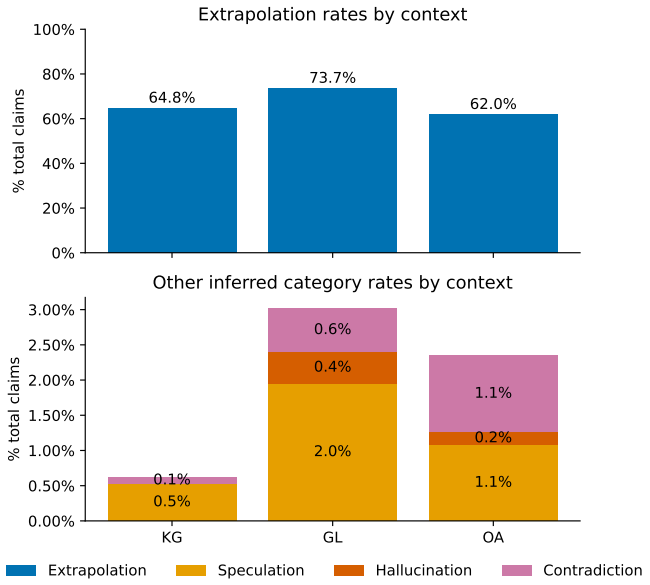

Figure 2. Inferred claim category composition per context (% of total claims).

Figure 2 visualizes these differences as stacked bars (% of total claims). The KG approach yields markedly fewer speculative and contradictory issues than both the GL and OA baselines, and reduces extrapolation by nine percentage points compared to GL. Despite these gains, many claims remain *inferred*, reflecting our design choice to allow clinically plausible, guideline-based extrapolations that may not be explicitly present in the patient record. These results support RQ1 (hallucination), RQ2 (contradiction), RQ3 (speculation), and RQ4 (extrapolation).

C. Qualitative Illustrations

Table V presents an excerpt of one narrative of the same case under the different grounding contexts. The *KG* narrative cites a unique guideline node *[27205d9]* and the recorded feature *RAS|wildtype*, both verifiable in the case file, demonstrating domain-rich yet fact-bound explanation. By comparison, the *GL* narrative, while fluent, infers "stage III disease" solely from *N1* and offers no patient-specific evidence for adjuvant need, showing readability at the expense of precision. The *OA* excerpt repeats guideline buzzwords ("high-risk stage III") relying on generic statements (*T3 N1 M0*), resulting in the most vague prose. For completeness, the last example in Table V presents an GL hallucination example. The mentioned fact—*"left-sided tumor (+0.04)"*—illustrates a feature not present in the patient case and most likely misattributed from the referenced guideline's (*§6.1.4.3.1.1*) metastatic EGFR-therapy discussion, underscoring how lack of authoritative grounding can introduce factual errors.

TABLE V. REPRESENTATIVE NARRATIVE EXCERPTS ACROSS GROUNDING CONTEXTS, WITH GL HALLUCINATION EXPLICITLY MARKED

Context	Narrative Excerpt
KG	Both guideline and model utilise an oxaliplatin + fluoropyrimidine backbone *[27205d9]*; the SHAP feature *RAS|wildtype* supports full cytotoxic sensitivity.
GL	The SHAP value for *N1 (0.28)* flags stage III disease and confirms the need for adjuvant therapy *(guideline §6.1.3)*.
OA	Both guideline and model emphasise high-risk stage III features *(T3 N1 M0)* as key drivers of therapy intensification.
GL	**Hallucination:** RAS wildtype (+0.03) and *left-sided tumor (+0.04)* slightly increased probability, mapping to metastatic guidelines for EGFR-directed therapy *(guideline §6.1.4.3.1.1)*.

The GL hallucination example highlights a reference to a non-existent feature (*left-sided tumor*).

Together, these qualitative vignettes also reinforce our quantitative results: The KG-grounded narrative delivers deep, context-rich explanations that remain verifiable, while the GL outputs sacrifice fidelity for readability and the OA outputs rely on overly generic statements, evidencing a tendency toward vagueness.

V. DISCUSSION

Our study demonstrates that anchoring narrative explanations in a guideline-derived KG improves factual reliability. The KG context reduced hallucinations to 0.0% of total claims in our sample—i.e., none were observed under this setup—supporting RQ1. Moreover, contradictions dropped to 0.1% and speculative claims to 0.5% of total claims, supporting RQ2 and RQ3 that KG grounding reduces both contradictions and speculation.

Moreover, anchoring explanations in the KG cut extrapolation rates from 73.7 % under the GL context to 64.8 %—a 9.0 percentage-point drop—demonstrating that guideline-derived KG grounding effectively constrains extrapolations to within established bounds and thereby confirms RQ4 (See Table IV).

Although the OA context exceeds KG in overall observed-claim rate (35.7% vs. 32.5%), its narrower input scope yields shallower, less semantically rich narratives. OA's lower extrapolation rate (61.9%) comes at the expense of actionable detail, whereas KG grounding delivers fully audit-ready, guideline-anchored explanations (See Table III and Figure 2). Finally, the relatively high share of *inferred* claims across conditions largely reflects clinically plausible, guideline-based extrapolations that provide useful framing but may not be directly present in patient records. In settings that require stricter evidencing, prompts or decoding constraints can restrict extrapolation at the cost of brevity; conversely, future work may calibrate this trade-off per user role (e.g., clinical vs. data science review).

These findings extend prior LLM explainers by showing that structured KG context not only enriches inference but also constrains factual drift [7]. We note that the absence of hallucinations should not be interpreted as impossibility; rather, it likely reflects the combination of KG constraints and the controlled, synthetic case distribution used here.

In practice, clinicians must rapidly validate AI recommendations. The traceable paths in KG narratives—linking each feature attribution to specific guideline nodes—can reduce expert review time by directly surfacing conflicts or affirmations in the guideline text. In our qualitative examples (Table V), KG narratives allowed unambiguous verification of treatment rationale, whereas GL outputs required additional cross-checking. We anticipate that integrating KG-grounded narratives into decision-support dashboards will shorten iteration cycles between data scientists and clinicians, as envisaged in collaborative AI workflows [25].

Our evaluation is constrained by some factors. First, we used 65 synthetic patient personas rather than real-world cases; while this allowed controlled variation, it may not capture the full complexity of clinical data. Second, we benchmarked against a single guideline (Onkopedia CRC) and one LLM version (GPT-o4-mini-high); generalization to other specialties or model variants remains to be demonstrated. Third, our error annotations—though 95% accurate in spot-checks—rely on an automated evaluation LLM; residual misclassifications could slightly bias absolute error rates. Finally, we measured only claim-level errors; additional dimensions such as usability, cognitive load, and end-user satisfaction were not assessed here.

VI. CONCLUSION AND FUTURE WORK

Having demonstrated through our evaluations that KG-grounded narrative explanations outperform both attribute-only and guideline-excerpt baselines in factual reliability, we now outline directions to build on this work. To address limitations and extend our findings, we propose the following directions: (1) Apply the pipeline to real-world data and diverse guidelines;

quantify clinician review time and simulated decision impact. (2) Iteratively refine KG-narrative prompts with user feedback and on-the-fly graph augmentation, aligning with human-centered XAI [5]. (3) Evaluate usability, trust calibration, and clinical actionability; extend metrics (e.g., comprehensiveness, empowerment).

Overall, our results confirm that fact-grounded narrative explanations built on guideline-derived Knowledge Graphs deliver superior factual reliability and coherence compared to attribute-only or guideline-excerpt baselines. By transparently linking model attributions to clinical evidence, this approach paves the way for more trustworthy, actionable AI in healthcare—bridging the critical gap between statistical performance and domain relevance.

ACKNOWLEDGMENT

This research has been funded by the German Federal Ministry of Education and Research (BMBF) under grant agreement no. 13FH5E11IA (CoHMed/NIO). Responsibility for the content of this publication lies with the author.

REFERENCES

[1] J. M. Duran and K. R. Jongsma, "Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai", *Journal of Medical Ethics*, vol. 47, no. 5, pp. 329–335, 2021. DOI: 10.1136/medethics-2020-106820.

[2] T. P. Quinn, S. Jacobs, M. Senadeera, V. Le, and S. Coghlan, "The three ghosts of medical ai: Can the black-box present deliver?", *Artificial Intelligence in Medicine*, vol. 124, p. 102 158, 2022, ISSN: 0933-3657. DOI: 10.1016/j.artmed.2021.102158.

[3] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions", in *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017.

[4] A. Bilal, D. Ebert, and B. Lin, "Llms for explainable ai: A comprehensive survey", *ACM Transactions on Intelligent Systems and Technology*, 2025, March 2025 edition.

[5] D. Martens, J. Hinns, C. Dams, M. Vergouwen, and T. Evgeniou, "Tell me a story! narrative-driven xai with large language models", *arXiv preprint*, 2023. eprint: 2309.17057.

[6] A. Zytek, S. Pido, S. Alnegheimish, L. Berti-Équille, and K. Veeramachaneni, "Explingo: Explaining ai predictions using large language models", in *IEEE Big Data Conference*, 2024. eprint: 2412.05145.

[7] J. Burton, N. A. Moubayed, and A. Enshaei, "Natural language explanations for machine-learning classification decisions", in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2023, pp. 1–9.

[8] M. A. Kadir, A. Mosavi, and D. Sonntag, "Evaluation metrics for xai: A review, taxonomy, and practical applications", in *27th IEEE International Conference on Intelligent Engineering Systems (INES)*, 2023, pp. 111–124. DOI: 10.1109/INES59282.2023.10297629.

[9] K. Matton, R. Ness, J. Guttag, and E. Kiciman, "Walk the talk? measuring the faithfulness of large language model explanations", in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[10] Y. Ming *et al.*, "Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"", *arXiv preprint*, 2024. eprint: 2410.03727.

[11] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, and H. Lakkaraju, "Are large language models post hoc explainers?", in *Robustness of Few-/Zero-Shot Learning Workshop @ NeurIPS 2023*, 2023.

[12] T. Lanham *et al.*, "Measuring faithfulness in chain-of-thought reasoning", *arXiv preprint*, 2023. eprint: 2307.13702.

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: A method for automatic evaluation of machine translation", in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[14] C. Lin, "Rouge: A package for automatic evaluation of summaries", in *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[15] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A comprehensive survey on automatic knowledge graph construction", *ACM Computing Surveys*, vol. 56, no. 4, 2024. DOI: 10.1145/3618295.

[16] F. Belleau, M. Nolin, N. Tourigny, A. Rigault, and J. Morissette, "Bio2rdf: Towards a mashup to build bioinformatics knowledge systems", *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 706–716, 2008. DOI: 10.1016/j.jbi.2008.03.004.

[17] Z. Chen, J. Chen, A. K. Singh, and M. Sra, "Xplainllm: A knowledge-augmented dataset for reliable grounded explanations in llms", in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore: Association for Computational Linguistics, 2024, pp. 7578–7596.

[18] Y. Gao *et al.*, "Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study", *JMIR AI*, vol. 4, e58670, 2025. DOI: 10.2196/58670.

[19] A. Khediri, H. Slimi, A. Yahiaoui, and M. Derdour, "Enhancing machine learning model interpretability in intrusion detection systems through shap explanations and llm-generated descriptions", in *6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, 2024. DOI: 10.1109/PAIS62114.2024.10541168.

[20] L. P. Meyer *et al.*, "Llm-assisted knowledge graph engineering: Experiments with chatgpt", in *First Working Conference on Artificial Intelligence Development for a Resilient and Sustainable Tomorrow (AIDRST 2023)*, C. Zinke-Wehlmann and J. Friedrich, Eds., ser. Informatik aktuell, Wiesbaden, Germany: Springer Vieweg, 2024. DOI: 10.1007/978-3-658-43705-3_8.

[21] F. Gaber *et al.*, "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis", *npj Digital Medicine*, vol. 8, p. 263, 2025. DOI: 10.1038/s41746-025-01684-1.

[22] Onkopedia Guidelines, "Colon carcinoma — onkopedia guideline", Accessed 2025-09, 2025, [Online]. Available: https://www.onkopedia.com/.

[23] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation", *Journal of Chronic Diseases*, vol. 40, no. 5, pp. 373–383, 1987. DOI: 10.1016/0021-9681(87)90171-8.

[24] R. G. Newcombe, "Two-sided confidence intervals for the single proportion: Comparison of seven methods", *Statistics in Medicine*, vol. 17, no. 8, pp. 857–872, 1998.

[25] M. Afshar, Y. Gao, D. Gupta, E. Croxford, and D. Demner-Fushman, "On the role of the umls in supporting diagnosis generation: Differential diagnoses proposed by large language models", *Journal of Biomedical Informatics*, vol. 157, p. 104 707, 2024. DOI: 10.1016/j.jbi.2024.104707.