A Taxonomy for Design and Evaluation of Prompt-Based Natural Language Explanations

Isar Nejadgholi [1] **Mona Omidyeganeh** [1] **Marc-Antoine Drouin** [1] **Jonathan Boisvert** [1]

Abstract

Effective AI governance requires structured approaches for stakeholders to access and verify AI system behavior. With the rise of large language models, Natural Language Explanations (NLEs) are now key to articulating model behavior, which necessitates a focused examination of their characteristics and governance implications. We draw on Explainable AI (XAI) literature to create an updated XAI taxonomy, adapted to prompt-based NLEs, across three dimensions: (1) *Context*, including task, data, audience, and goals; (2) *Generation and Presentation*, covering generation methods, inputs, interactivity, outputs, and forms; and (3) *Evaluation*, focusing on content, presentation, and user-centered properties, as well as the setting of the evaluation. This taxonomy provides a framework for researchers, auditors, and policymakers to characterize, design, and enhance NLEs for transparent AI systems.

1. Introduction

As AI is increasingly integrated into critical domains such as healthcare, finance, law, and defense, developing frameworks that bridge the gap between technical capabilities and governance requirements becomes essential (Reuel et al., 2024). The black-box nature of many AI models raises significant concerns around fairness, trust, and accountability (Goodman & Flaxman, 2017), and poses a barrier to effective governance by limiting access, and verification mechanisms. Explainable AI (XAI) directly addresses this opacity by providing methods and strategies that enhance transparency and make the decision-making processes of AI systems more understandable to humans through the use of *explanations* (Das & Rad, 2020). In general, explanations offer insights into how the AI system arrives at a specific decision and aim to empower certain stakeholders to *access* and *verify* the behavior of these systems effectively.

Traditionally, XAI has been more focused on making the inner workings of deep learning models understandable to AI developers, with the goal of debugging and improving such systems (Minh et al., 2022). However, most of these explanations were not accessible to other users, auditors or policymakers. With the emergence of LLMs, a unique opportunity arose in the field of XAI. Now, LLMs can be directly prompted and integrated into AI systems to provide textual explanations which are referred to as prompt-based Natural Language Explanations (NLEs) (Costa et al., 2018). Such explanations serve as crucial interfaces between AI systems and human stakeholders due to their unprecedented fluency (Lakkaraju et al., 2022).

While NLEs offer unique opportunities for enhancing AI transparency, they unfortunately introduce new challenges for governance due to their potential to present plausible yet unverified or misleading assertions. Despite many studies demonstrating successful human-AI collaboration through incorporating NLEs (Wiegreffe et al., 2022), a growing body of research has shown that, if not carefully designed, these explanations can reinforce poor decisions. The limitations of NLEs include inconsistent or nonfactual explanations (Ye & Durrett, 2022), poor reasoning abilities (Turpin et al., 2023) increase of over-reliance on the AI system (Zhang et al., 2020; Wang & Yin, 2021; Poursabzi-Sangdeh et al., 2021; Bansal et al., 2021; Chen et al., 2023), and misleading outputs (Kayser et al., 2024). This evidence calls for standardizing such explanations and creating guidelines that different stakeholders can adhere to when designing and validating such explanations.

Specifically, the growing use of prompt-based NLEs necessitates a dedicated taxonomy that captures their distinct characteristics and governance implications. While the field of XAI has produced numerous taxonomies (Arrieta et al., 2020; Speith, 2022), none, to the best of our knowledge, are specifically tailored to NLEs. This work addresses that gap by adapting existing XAI taxonomies to the context of prompt-based NLEs, with the goal of informing their design and evaluation. Such a taxonomy advances technical AI governance by enabling the verification of model claims, providing meaningful access to system behavior for diverse

[1]National Research Council Canada, Canada. Correspondence to: Isar Nejadgholi <isar.nejadgholi@nrc-cnrc.gc.ca>.

Workshop on Technical AI Governance (TAIG) at ICML 2025, Vancouver, Canada.

stakeholders, and supporting the practical implementation of core principles such as accountability and transparency.

2. Methodology and Scope

Our work builds on the XAI concepts synthesized in a recent meta-analysis by Schwalbe & Finzel (2024), which consolidates insights from over 70 surveys into a unified taxonomy of methods, traits, and evaluation metrics. We refer to this tool as the XAI meta-taxonomy and adapt it to the specific case of NLEs. A visual representation of the XAI meta-taxonomy is shown in Figure A.1, Appendix A. For adaptation, we incorporate NLE-specific considerations from relevant literature, such as Cambria et al. (2023), who proposed a roadmap to understanding NLEs. We present our arguments for this adaptation and the related literature in section 3. Here, we clarify the definition of NLEs, the scope of our study and our adaptation approach.

Definition of NLEs: Inspired by Cambria et al. (2023) and Ribeiro et al. (2016), we define NLEs as textual artifacts that **convey** a qualitative understanding of how an **input's components** influence a model's prediction, along with **contextual information** relevant to a specific user or the task in general. This definition distinguishes NLEs from raw model reasoning traces or other outputs that are not centred explicitly around communication with the user.

NLEs within the XAI landscape: Schwalbe & Finzel (2024) synthesize the XAI literature and present relevant concepts under three high-level procedural components: *Problem Definition*, *Explanator*, and *Metrics*, with multiple hierarchical sub categorizations. While these components remain highly relevant for NLEs in general, their detailed subcategories require adaptation for the case of NLEs. Some elements are less directly applicable (e.g., formal mathematical constraints). Other subcategories demand greater nuance (e.g., interactivity in dialogue-based systems) or require more specificity in definitions.

First, while in general, a generative model might be fine-tuned or adapted otherwise to generate specialized explanations (Sammani et al., 2022), we limit the scope of this study to *prompt-based* NLEs. Next, we narrow the scope of our work to a common class of prompt-based NLEs that are: (1) *post-hoc* – generated after the model has produced a decision (Retzlaff et al., 2024); (2) *model-agnostic* – not reliant on the internal architecture or parameters of the model (Ribeiro et al., 2018); and (3) *local* – produced with respect to individual instances rather than the model as a whole (Lundberg et al., 2020). Within this scope, we exclude components of XAI that pertain to inherently interpretable, self-explaining models or model-specific explanation techniques (Rudin, 2019), and *global* explanations that describe a model's overall behavior (Saleem et al., 2022).

Methodological approach: Our adaptation of the XAI meta-taxonomy was informed by an iterative, consensus-driven process involving experts from multiple domains, including the authors. Specifically, we engaged stakeholders, including a researcher in explainable AI, several model developers, a technical manager, and an organizational decision-maker. Through a series of collaborative reviews, we identified which components of the original taxonomy align with the properties of prompt-based NLEs and where extensions or refinements were necessary.

3. Proposed Taxonomy of Prompt-Based NLE

Table 1 presents our proposed taxonomy for post-hoc, model-agnostic, and local prompt-based NLEs, adapted from the meta-taxonomy (Figure A.1). We elaborate on each component and its nested subcategories within a two-tier structure.

3.1. Context Definition

Context Definition is the first component of designing NLEs. This element corresponds to the *Problem Definition* phase of the XAI meat-taxonomy, which aims to formulate the requirements of the task and the use case aspect. The subcategories of *Task Type* and *Data Type* remain critically relevant to the case of NLEs. The former recognizes that the nature of the explanation depends on the underlying AI task, such as classification (explaining category assignment), Regression (justifying numerical predictions), or other tasks such as detection. The latter differentiates between explanations for models trained on structured data (e.g., tabular data with defined variables) versus unstructured data (e.g., text, images, graphs), acknowledging that the form and content of explanations may differ significantly. Explanations for graph neural networks, for instance, present unique challenges (Yuan et al., 2022).

As explained in Section 2, we omit the interpretability subcategory that pertains to explaining the intertwined process of prediction and explanation when they are performed jointly (Du et al., 2019). This exclusion is justified by the fact that NLEs are typically employed as *post-hoc* methods, i.e., they are generated *after* the model has made a decision.

Under the *Context Definition*, we add a new subcategory, *Audience*, which is crucial for NLE but not well-represented in the XAI meta-taxonomy. Due to the versatility of NLEs, they can be generated for different stakeholders, and therefore, considering the audience is a major aspect of the *context definition* for NLE. Here, we consider the following five categories of agents, proposed by (Tomsett et al., 2018), who might be the audience of NLE: **1) Creator**: Agents who create the system, including owners and developers, and might need NLE for debugging, error analysis, etc.,

Table 1. Taxonomy for post-hoc, model-agnostic, and local prompt-based NLEs, adapted from existing XAI literature.

Component	Subcategory	Description/Example
Context Definition	Task Types	Classification, regression, clustering, generation, etc.
	Data Types	Structured (e.g., tables), unstructured (e.g., text, images, graphs).
	Audience	Creator, Operator, Executor, Decision Subject, Examiner
	Explanation Goal	Model reasoning, task-level justification, or error detection
Generation & Presentation	Model Type	Generative LLMs, VLMs.
	Input	Prompt and context artifacts provided by users.
	Interactivity	Dialogue-based refinement of explanation scope and relevance.
	Output Type	Textual formats: examples, counterfactuals, structured narratives.
	Presentation Form	Plain text, visuals, interactive widgets; adapted to user expertise and needs.
Evaluation	Content Properties	Correctness, Completeness, Consistency, Continuity, Contrastivity, Comprehensibility
	Presentation Properties	Compactness, Composition, Confidence, Translucence
	User-Centered Properties	Actionability, Personalization, Coherence, Controllability, Novelty
	Evaluation setting	Functionally Grounded (automated), Human-Grounded (subjective ratings), and Application-Grounded (performance in real-world tasks).

2) Operator: Agents who directly interact with the AI by providing the system with inputs and directly receiving the system's outputs and might be interested in simple and informative NLEs to interpret the output or detect the model's failures. **3) Executors**: Agents who make decisions based on the output of the system and might be interested in contextual information relevant to the decision they want to make. Domain-specific jargon might be acceptable or even preferred. **4) Decision Subject**: Agents whose data is used in the training of the model, commonly interested in knowing how the system is using their data. **5) Examiners**: Agents who audit or investigate the model, potentially interested in explanations that describe why some rules or regulations have been violated or complied with.

Moreover, we added the subcategory of *Explanation Goal* to the *context definition* category as it can vary based on the use case and audience. For this, we adopt the golas of NLE generation proposed by (Chen et al., 2022), including: **1) Human Understanding of the AI Model:** Explain the model's decision boundary so that humans can understand the model's reasoning. **2) Informing Human About the Task:** Explain the decision boundary of the task itself to enable humans to make better decisions and **3) Enabling Users to Detect Model Errors:** Provide contextual, counterfactual, or confidence-level information to help users identify potential model errors.

3.2. Generation and Presentation

We rename the *Explanator* module of the XAI meta-taxonomy to *Generation and Presentation* of explanations in Table 1. This component concerns how the explanation is generated and presented to users. Within the subcategories of this component, mathematical constraints like linearity or monotonicity are less critical for NLEs, as explanations are delivered in natural language rather than structured mathematical formats (e.g., decision rules or equations). For example, while a decision tree might require sparsity constraints, an NLE is usually agnostic to such constraints. We also omit the *portability* and *locality* properties as we are focused on specific cases of model-agnostic and local (individual instance level) explanations.

Determining the *explanator model type*, i.e., the model that is prompted to generate the explanations, is the first step. This can be one of the large families of generative models, including large language models (Kumar, 2024) or vision language models (Zhang et al., 2024). *Input* to the explanator is another subcategory where the designer of NLE identifies the user prompt and other artifacts that are fed to the generative model. Furthermore, *interactivity* is a lot more feasible in the case of NLE compared to traditional XAI models, as users can engage in dynamic dialogues to refine explanations. However, the line between *interactivity* and required *input*, is blurred in the case of NLE since user feedback, data, and context can all be provided through user interactions in the form of prompts. Incorporating insights from the field of human-computer interaction is essential in optimizing the interplay between interactivity and input types in prompt-based NLE (Reddy, 2024).

Turning to the *output type* subcategory, NLE outputs can range from concise summaries to structured narratives, depending on context. This depends on the explanation goal and the desired NLE's characteristics (see section 3.3).

For the *presentation* subcategory, the presentation of NLEs is highly adaptable and can be tailored to the user's expertise and goals. For example, technical users may receive explanations with statistical terms (e.g., the confidence score), and laypersons might prefer simplified analogies (e.g., flagged transaction to spot unusual activity). Also, while the text is the primary presentation format, NLEs can integrate visual aids such as highlighting parts of input images (Kayser et al., 2024), or interactive widgets such as sliders to explore "what-if" scenarios (Meyer & Zhu, 2024).

3.3. Evaluation

This component evaluates the quality of explanations (regardless of the predictive performance). Evaluating explanations is inherently complex due to the subjectivity of what a "good" explanation is, its dependence on context, the absence of a universally accepted ground truth, and the many desirable characteristics of an ideal explanation. Also, in XAI, a central tension exists between faithfulness (accuracy in reflecting the reasoning) and plausibility (being perceived as reasonable or satisfying by humans) (Jacovi & Goldberg, 2020). Prompt-based NLEs tend to favor plausibility over strict faithfulness, which comes at the cost of potentially oversimplifying or misrepresenting model internals.

In order to identify the potential desired characteristics of NLEs, we start with a framework with twelve explanation properties, Co-12, originally proposed by Nauta et al. (2023). These properties are organized under three categories *Content*, *Presentation*, and *User-centered* properties. The *Content* category comprises: 1) Correctness, 2) Completeness, 3) Consistency, 4) Continuity, 5) Contrastivity, and 6) Covariate Complexity, and mainly focuses on the faithfulness of the explanation with respect to the model or task. *Presentation* properties include: 7) Compactness, 8) Composition, 9) Confidence concern how explanations are formatted and structured. Finally, *User-centered* properties, include: 10) Context, 11) Coherence, and 12) Controllability, and emphasize the relevance and utility of explanations for end users. *Presentation* and *User-centered* properties generally fall under the plausibility criteria.

Table 2 presents 15 potential characteristics of NLEs, along with their descriptions, adapted from the above-mentioned Co-12 properties. Leveraging the usage-context-driven evaluation framework of Liao et al. (2022), we adapt the Co-12 framework to capture the unique affordances of NLEs. Specifically, 1) We rename *Covariate Complexity* as *Comprehensibility*, aligning with the terminology of Liao et al. (2022), which is more intuitive for language-based explanations. 2) We decompose the original *Context* property (under the *User-centered dimension*) into two distinct criteria, *Personalization* (tailoring to user needs) and *Actionability* (supporting decision-making), given the increased contextual adaptability of NLEs. 3) We add *Translucence* to the *Presentation* category to reflect an NLE's ability to communicate its own limitations or conditions of validity. 4) We include *Novelty* under *User-centered* properties, recognizing that NLEs can provide new or surprising information derived from pre-trained knowledge.

These properties might be evaluated in three settings: *Application Grounded*, *Human-Grounded* and *Functionally Grounded* evaluations (Doshi-Velez & Kim, 2017). Application-grounded evaluation credits explanations that improve the task in the real-world application and tests the explanations in the application context. While reaching high application-grounded results is the ultimate goal in generating explanations, designing and conducting such evaluations is resource-intensive and requires highly skilled users. Human-grounded evaluations offer an intermediate way of anticipating how the explanations would work in the real world by running human studies with laypeople. While human-grounded evaluations offer valuable insights into human perception, they can still be time-consuming. Functionally grounded evaluations provide alternative approaches that can be automated and do not require human.

Table 2. Potential Desirable Explanation Properties

	Property	Description
Content	Correctness	Accurately reflects AI's reasoning.
	Completeness	Covers all relevant factors contributing to the output.
	Consistency	Produces identical explanations for identical inputs.
	Continuity	Small input change results in small explanation change.
	Contrastivity	Highlights differences from alternative outcomes.
	Comprehensibility	Uses human-understandable concepts and relations.
Presentation	Compactness	Provides succinct and non-redundant explanations.
	Composition	Presents information clearly and structurally.
	Confidence	Communicates the model's certainty or uncertainty.
	Translucence	Acknowledges explanation limitations or alternatives.
User-centered	Actionability	Supports the user in deciding what to do next.
	Personalization	Tailored to the user's context or needs.
	Coherence	Aligns with user expectations and prior knowledge.
	Controllability	Enables user interaction or refinement of explanations.
	Novelty	Provides new or non-obvious information to the user.

4. Example Use Case

In this section, we elaborate on the application of the proposed taxonomy in "anomaly detection in remote sensing using airborne imagery". This task is crucial for public safety, border security, intelligence, surveillance, and reconnaissance, as well as natural disaster response. Anomalies include unusual or unexpected events, which signal important safety risks or security threats, such as "a bus travelling on a bike path" or a "person standing in the middle of a roadway". An effective AI-based system is not only expected to accurately detect such abnormalities but also to explain their nuances, to enable better-informed decision-making.

Figure B.1 highlights an example where a vision-language-based system is deployed to detect anomalies from images taken by a drone and communicate the necessary information to an operator. In this setting, the operator requires clear and reliable explanations that justify why the system identified a particular behavior as abnormal. The taxonomy supports this need by setting the *Explanation Goal* to clarify "*why the behavior deviates from normal patterns*". It also guides the system to present the explanation in a structured format including 1) a binary decision (anomaly or not), 2) contrast between expected and observed behavior, and 3) a confidence level. This systematized prompting makes the ex-

planation more straightforward to interpret, use and evaluate. For evaluation, Correctness, Confidence and Actionability of NLEs can be assessed in an application-grounded setting.

The taxonomy also recognizes that different users need different explanations: For developers, it is crucial to identify errors, such as false alarms, and analyze the model's decision-making process for debugging purposes. Therefore, for this audience the goal of explanations shifts from "*Why the behavior deviates from normal patterns?*" (as it is for the operator) to "*Why/how does the model detect this behavior as an anomaly?"*. The explanation may be presented as highlighted regions in the image or as counterfactual features showing what would change the model's decision.

On the other hand, executors, such as field agents or decision-makers, need contextual and actionable explanations to understand the anomaly. Therefore, for them, the goal of explanations shifts to *"What is the situation, and how should I respond?"* The response may include concise summaries of the anomaly, its potential implications, and recommended next steps to enable rapid and effective action in the field. Finally, examiners, such as auditors or policy reviewers, need explanations that demonstrate the system's compliance with regulations, ethical standards, and operational guidelines. For examiners, the explanation goal becomes *"Does this decision align with policy and was it reached in a valid way?"*. The response may take the form of a justification that includes traceable reasoning steps, references to decision criteria, and documentation of relevant inputs to support regulatory review and accountability.

5. Discussion

While the presented taxonomy provides a general framework for designing prompt-based NLEs, the most relevant elements should be empirically identified for each task with the following critical considerations in mind.

Faithfulness as a necessity: Although we mostly focused on the communicative aspect of NLEs, we recognize that, regardless of audience or context, explanations that distort, oversimplify, or obscure the system's actual workings risk undermining transparency, accountability, and trust. Effective explanation systems must therefore not only adapt to the user and context, but also remain internally consistent and grounded in the model's underlying principles. In some cases, selective disclosure may be appropriate to accommodate user goals or protect sensitive information, but this must not compromise factual accuracy. Guaranteeing faithfulness is a challenging technical task that is beyond the scope of this study and deserves special attention (Yeh et al., 2019; Lyu et al., 2024).

Subjectivity of Criteria: Identifying an optimum prompt requires a comparative analysis of several settings based on the taxonomy, with respect to the desired criteria. However, many of the desired criteria we mentioned are highly subjective, which makes their evaluation particularly challenging. One effective paradigm in early design stages is to use the LLM-as-a-judge testing (Gu et al., 2024), for automatic and cost-effective preliminary assessment. This approach involves using more advanced LLMs to evaluate the characteristics of the explanations generated by a production-grade model, and is suitable for early refinements before investing in high-cost, multi-annotator evaluations. However, care should be taken in such evaluations due to the limitations of LLMs when deployed as judges (Szymanski et al., 2025).

Trade-offs and conflicts Crucially, there is a fundamental tension between some of the desired characteristics of NLEs shown in table 2. For example, achieving perfect *Correctness* might result in an explanation that is highly complex and detailed. Such an explanation, while greatly faithful and informative, could be difficult for a non-expert user to understand, and scores low on *Comprehensibility*. The ideal balance between these properties often depends on the specific application context and the needs of the user, especially in high-stakes, time-sensitive settings.

Specifically, AI assistance in sensitive or time-pressured environments may pose conflicts and risks that require striking a critical balance. For example, Swaroop et al. (2024) found trade-offs between decision accuracy and speed, where simpler explanations led to faster decisions but increased the risk of overreliance on the AI, while more detailed explanations improved understanding but slowed down the response. These effects are further influenced by individual user traits, such as their baseline trust in AI and need to be mitigated relative to context, user goals, and cognitive constraints.

6. Conclusion

Our taxonomy brings together key concepts relevant to NLEs and offers a practical checklist for designers and evaluators while supporting a systematic approach to AI governance. It aligns with several capacities outlined by Reuel et al. (2024), including: access, by enabling interaction through natural language without needing internal model access; and verification, by guiding the design of explanations that can substantiate claims about system behavior. Moreover, this taxonomy could support the operationalization of transparency and accountability by aligning explanation strategies with specific audiences and objectives. Future work involves validating its use through experiments in collaboration with human-computer interaction experts.

Acknowledgment

This work was conducted by the National Research Council Canada on behalf of the Canadian AI Safety Institute.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16, 2021.

Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., and Nobani, N. A survey on xai and natural language explanations. *Information Processing & Management*, 60 (1):103111, 2023.

Chen, C., Feng, S., Sharma, A., and Tan, C. Machine explanations and human understanding. *arXiv preprint arXiv:2202.04092*, 2022.

Chen, V., Liao, Q. V., Wortman Vaughan, J., and Bansal, G. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7 (CSCW2):1–32, 2023.

Costa, F., Ouyang, S., Dolog, P., and Lawlor, A. Automatic generation of natural language explanations. In *Companion Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pp. 1–2, 2018.

Das, A. and Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Du, M., Liu, N., and Hu, X. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1): 68–77, 2019.

Goodman, B. and Flaxman, S. European union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine*, 38(3):50–57, 2017. doi: 10.1609/aimag.v38i3.2741. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2741.

Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Jacovi, A. and Goldberg, Y. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Kayser, M. G., Menzat, B., Emde, C., Bercean, B. A., Novak, A., Morgado, A. T. E., Papiez, B., Gaube, S., Lukasiewicz, T., and Camburu, O.-M. Fool me once? contrasting textual and visual explanations in a clinical decision-support setting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18891–18919. Association for Computational Linguistics, 2024.

Kumar, P. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260, 2024.

Lakkaraju, H., Slack, D., Chen, Y., Tan, C., and Singh, S. Rethinking explainability as a dialogue: A practitioner's perspective. In *NeurIPS Workshop on Human Centered AI*, 2022.

Liao, Q. V., Zhang, Y., Luss, R., Doshi-Velez, F., and Dhurandhar, A. Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable ai. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pp. 147–159, 2022.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67, 2020.

Lyu, Q., Apidianaki, M., and Callison-Burch, C. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pp. 1–67, 2024.

Meyer, L. S. and Zhu, J. Slide to explore'what if': An analysis of explainable interfaces. In *Adjunct Proceedings of the 2024 Nordic Conference on Human-Computer Interaction*, pp. 1–6, 2024.

Minh, D., Wang, H. X., Li, Y. F., and Nguyen, T. N. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, pp. 1–66, 2022.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation

methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–52, 2021.

Reddy, S. T. A. Human-computer interaction techniques for explainable artificial intelligence systems. *Recent Trends in Artificial Intelligence & It's Applications*, 3:1–7, 2024.

Retzlaff, C. O., Angerschmid, A., Saranti, A., Schneeberger, D., Roettger, R., Mueller, H., and Holzinger, A. Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists. *Cognitive Systems Research*, 86:101243, 2024.

Reuel, A., Bucknall, B., Casper, S., Fist, T., Soder, L., Aarne, O., Hammond, L., Ibrahim, L., Chan, A., Wills, P., et al. Open problems in technical ai governance. *arXiv preprint arXiv:2407.14981*, 2024.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Saleem, R., Yuan, B., Kurugollu, F., Anjum, A., and Liu, L. Explaining deep neural networks: A survey on the global interpretation methods. *Neurocomputing*, 513:165–180, 2022.

Sammani, F., Mukherjee, T., and Deligiannis, N. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8322–8332, 2022.

Schwalbe, G. and Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5):3043–3101, 2024.

Speith, T. A review of taxonomies of explainable artificial intelligence (xai) methods. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 2239–2250, 2022.

Swaroop, S., Buçinca, Z., Gajos, K. Z., and Doshi-Velez, F. Accuracy-time tradeoffs in ai-assisted decision making under time pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 138–154, 2024.

Szymanski, A., Ziems, N., Eicher-Miller, H. A., Li, T. J.-J., Jiang, M., and Metoyer, R. A. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 952–966, 2025.

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, 2018.

Turpin, M., Michael, J., Perez, E., and Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023.

Wang, X. and Yin, M. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 318–328, 2021.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., and Choi, Y. Reframing human-AI collaboration for generating free-text explanations. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 632–658, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.47. URL https://aclanthology.org/2022.naacl-main.47/.

Ye, X. and Durrett, G. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392, 2022.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

Yuan, H., Yu, H., Gui, S., and Ji, S. Explainability in graph neural networks: A taxonomic survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5782–5799, 2022.

Zhang, J., Huang, J., Jin, S., and Lu, S. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 295–305, 2020.

A. Meta-Taxonomy of XAI

Figure A.1 presents the meta-taxonomy, whose components we systematically analyzed to develop our proposed taxonomy.

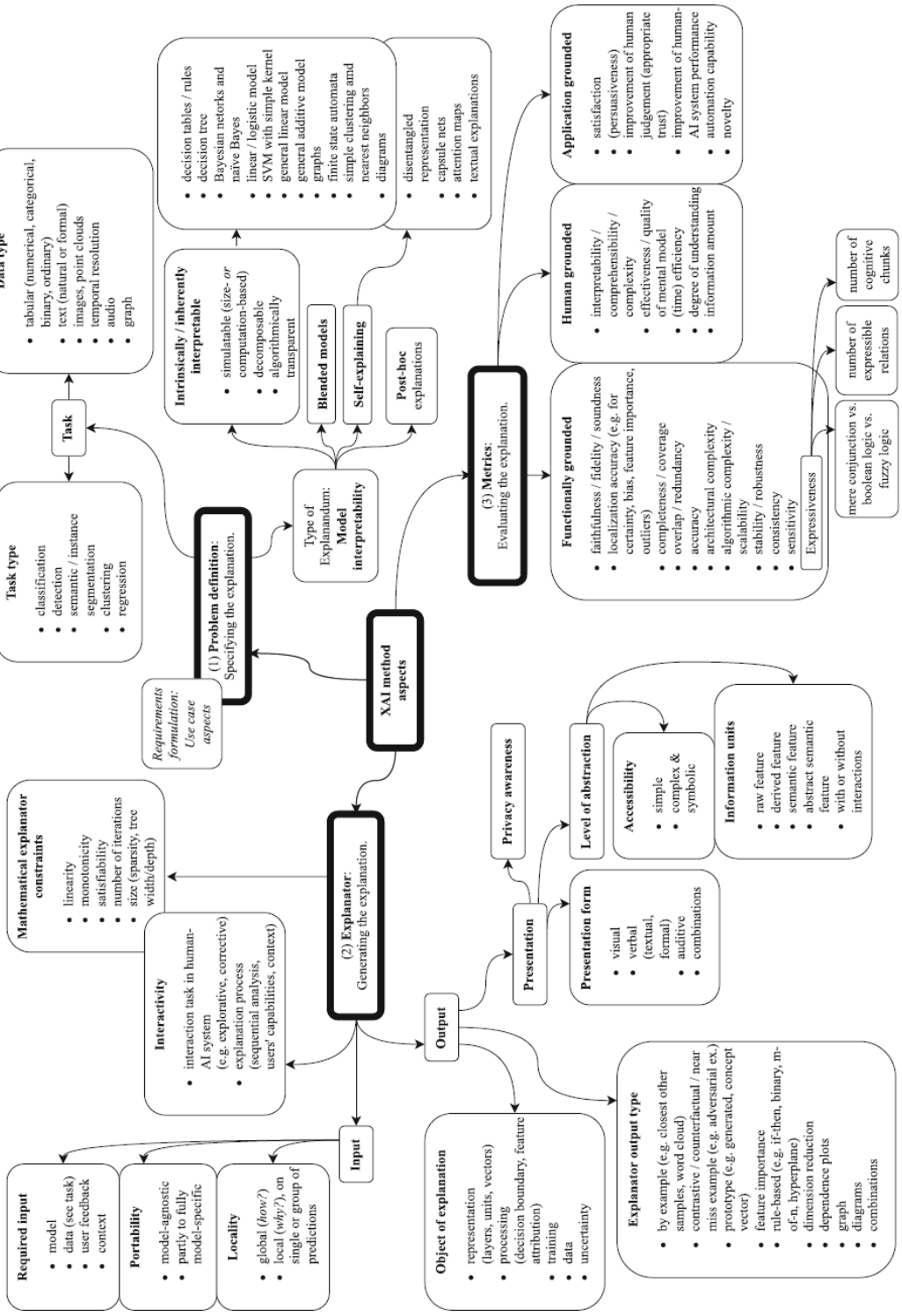
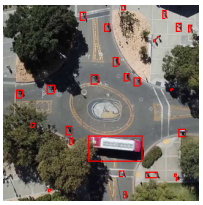


Figure A.1. The XAI meta-taxonomy proposed by Schwalbe & Finzel (2024) (Page 50, Figure 7).

B. Example Prompt Guided by Taxonomy

Here, we show an example explanation for the task of traffic anomaly detection guided by the taxonomy.



SYSTEM PROMPT

TASK TYPE:
Your task is to detect and explain anomalies in the provided input. You carefully analyze the input image to detect irregularities or deviations from expected patterns or behaviors within the given image.

DATA TYPE:
You work with images captured by a camera mounted on an aerial vehicle. Red boxes in the image represent moving objects.

AUDIENCE:
Your explanations will be used by a technical operator who manages the aerial vehicle.

EXPLANATION GOAL:
In your explanation, you provide evidence for **WHY** the detected behaviour is an anomaly. Human operator uses your explanation to understand the nature of the anomaly.

OUTPUT TYPE:
You generate clear, precise textual explanations that include:

- Binary anomaly detection decision
- Description of both expected behaviors and observed deviations with contrastive argument.
- A binary confidence level (high/low) for the identified anomaly.

PRESENTATION FORMAT:

- **Anomaly Detected:** This only includes YES or NO. YES refers to the presence of anomalies, and NO refers to normal cases. Do not include anything else.
- **Expected Normal:** This includes a detailed description of typical behavior or pattern expected in the region.
- **Observed Anomaly:** This includes a clear contrast between the observed deviation and the expected behavior.
- **Confidence Level:** This states the confidence level regarding the presence of the anomaly in the image and includes ”High” or ”LOW, Manual Check Recommended”. Do not include anything else.

RESPONSE

Anomaly Detected: Yes

Expected Normal: In a typical roundabout, vehicles are expected to navigate around the central island in a counter-clockwise direction, following the designated lanes and road markings. The roundabout is designed to facilitate smooth traffic flow, with vehicles entering and exiting at designated points without obstruction.

Observed Anomaly: The image shows a bus positioned across the roundabout, blocking the flow of traffic. This is a deviation from the expected behavior as the bus is not following the roundabout’s circular path and is instead obstructing the lanes, potentially causing a traffic jam and disrupting the normal flow of vehicles.

Confidence Level: High

Figure B.1. A system prompt for the task of traffic anomaly detection guided by the XAI taxonomy proposed in Table 1 and the generated response by gpt-4o-2024-08-06 for the image shown above (without user prompt). Red boxes within the image represent moving objects. The image is sourced from `https://gitlab.lrz.de/ai4eo/reasoning/drone-anomaly` and red boxes are added through implementation of computer vision techniques. This dataset is originally used to evaluate existing anomaly detection methods and provides a benchmark for the task of training and testing video processing models. The authors would like to acknowledge Joshua Power and Delasie Torkornoo, Software Developer Specialists at NRC, for their contribution to producing the results presented in this appendix.