Explainable Qualitative Analysis with Large Language Models

Asma Maheen

University of Gujrat, Pakistan

Corresponding Email: asmamaheen410@gmail.com

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in reasoning, summarization, and natural language understanding across diverse domains. However, despite their growing adoption in decision-critical settings, the interpretability of their qualitative judgments remains limited. This paper explores explainable qualitative analysis using LLMs, focusing on how such models generate, structure, and justify non-numerical insights. We argue that qualitative reasoning—such as thematic interpretation, sentiment justification, and contextual inference—requires distinct explainability mechanisms beyond traditional feature attribution. A structured framework is proposed to extract, analyze, and validate explanations generated by LLMs during qualitative tasks. Through controlled experiments on text interpretation and expert-aligned reasoning benchmarks, we evaluate the faithfulness, consistency, and human-alignment of LLM-generated explanations. The findings highlight both the promise and current limitations of LLMs as explainable qualitative analysts and provide practical insights for deploying them responsibly in research and decision-support systems.

Keywords: Large Language Models, Explainable AI, Qualitative Analysis, Interpretability, Natural Language Reasoning, Human-Centered AI

I. Introduction

The rise of Large Language Models has reshaped how machines interact with human language, enabling systems to perform tasks that were traditionally considered subjective or qualitative in nature [1]. These tasks include interpreting narratives, extracting themes, analyzing opinions, and generating explanations that resemble human reasoning. While quantitative prediction has long been the focus of machine learning research, qualitative

analysis presents a different challenge—one that requires models not only to produce outputs, but to justify them in ways humans can understand and trust.

In many real-world applications such as social science research, policy analysis, healthcare narratives, and legal reasoning, decisions are driven by qualitative judgments rather than numerical scores. In these contexts, the value of an LLM lies not merely in what conclusion it reaches, but in how it explains that conclusion [2]. Without transparency into the reasoning process, even highly accurate qualitative outputs may fail to gain acceptance among domain experts and stakeholders.

Explainable Artificial Intelligence has traditionally focused on interpreting numerical models through feature importance, saliency maps, or surrogate models [3]. However, these techniques are poorly suited for LLMs operating on high-dimensional linguistic representations. Qualitative reasoning emerges from distributed semantic patterns, implicit world knowledge, and contextual inference, making conventional explainability tools insufficient for understanding LLM behavior in such settings [4].

This paper argues that explainability for qualitative analysis must be reframed. Instead of asking which input tokens caused an output, we must examine how models structure arguments, select evidence, and maintain internal consistency across explanations. Explainability, in this sense, becomes an analysis of reasoning narratives rather than numerical contributions.

The objective of this work is to systematically study explainable qualitative analysis with LLMs. We seek to understand what kinds of explanations LLMs generate, how reliable those explanations are, and how closely they align with human expert reasoning. By grounding this investigation in experimental evaluation, we aim to move beyond anecdotal observations toward measurable insights.

II. Related Work

Prior research on explainable AI has largely concentrated on supervised learning models, particularly in vision and tabular domains [5]. Techniques such as LIME, SHAP, and attention visualization have been widely adopted to provide post-hoc explanations. While

these methods offer valuable insights for structured data, they struggle to capture the abstract and contextual nature of language-based reasoning performed by LLMs.

Recent studies have explored attention mechanisms as a proxy for explanation in transformer models. However, it has been shown that attention weights do not reliably correspond to causal reasoning, especially in deep language models. As a result, attention-based explanations often provide an illusion of transparency rather than a faithful account of model behavior, particularly in qualitative tasks involving long-form reasoning.

Another line of work focuses on chain-of-thought prompting, where models are encouraged to produce intermediate reasoning steps. While this approach improves performance and interpretability, it raises concerns about whether the generated explanations truly reflect internal reasoning or are simply plausible post-hoc narratives [6]. This distinction is critical when LLMs are used as qualitative analysts rather than answer generators.

Qualitative analysis in computational social science and digital humanities has traditionally relied on rule-based systems or topic models. Although these approaches are interpretable, they lack the flexibility and contextual understanding of modern LLMs. The trade-off between interpretability and expressiveness has thus remained a persistent challenge.

More recently, research has begun to examine explanation faithfulness and consistency in LLMs. These studies suggest that while LLMs can generate convincing explanations, their justifications may vary significantly across prompts or contradict earlier reasoning. This paper builds upon these findings by explicitly evaluating explainability in the context of qualitative analysis tasks [7].

III. Methodology

The proposed methodology treats explainable qualitative analysis as a structured interaction between input text, model reasoning, and generated explanation. Rather than assuming explanations are inherently meaningful, we analyze them as artifacts that can be evaluated for coherence, evidence usage, and alignment with expert reasoning [8]. This approach allows us to systematically study explanation quality without assuming internal model transparency.

We define qualitative analysis tasks as those requiring interpretation rather than prediction. Examples include identifying dominant themes in narratives, explaining sentiment polarity with justification, and interpreting implicit intent. For each task, the model is prompted to provide both an answer and an accompanying explanation that references elements of the input text.

To ensure consistency, we adopt a standardized prompting strategy that requests explicit reasoning steps without imposing rigid templates. This balances natural language freedom with analytical structure, allowing explanations to emerge organically while remaining evaluable. Multiple prompt variants are used to test robustness and explanation stability.

The evaluation framework focuses on three dimensions: faithfulness, consistency, and human alignment. Faithfulness measures whether explanations are grounded in the input text rather than hallucinated content [9]. Consistency evaluates whether explanations remain stable across semantically equivalent prompts. Human alignment assesses similarity between model explanations and expert-generated qualitative justifications.

Expert annotations are collected from domain specialists who independently analyze the same texts and provide written explanations. These expert responses serve as a reference point, not as ground truth, recognizing that qualitative reasoning inherently allows for multiple valid interpretations.

IV. Experimental Setup

The experimental evaluation is conducted on a curated dataset consisting of opinion articles, interview transcripts, and policy documents. These texts are chosen because they require nuanced interpretation rather than factual extraction. Each document is annotated by experts for thematic focus, sentiment, and interpretive rationale.

We evaluate multiple state-of-the-art LLMs under identical prompting conditions to ensure comparability [10]. Models are instructed to perform qualitative analysis and generate explanations in complete sentences, explicitly referencing textual evidence where possible. Temperature and decoding parameters are fixed to minimize variability due to sampling randomness.

To assess faithfulness, we employ a textual grounding metric that measures overlap between explanation content and source text. Explanations that introduce unsupported claims or external facts are penalized. This helps identify cases where models produce persuasive but ungrounded reasoning.

Consistency is evaluated by rephrasing prompts while keeping the underlying task unchanged. Explanation similarity is measured using semantic similarity scores and qualitative comparison [11]. Large deviations in reasoning structure indicate instability in the model's explanatory behavior.

Human alignment is assessed through blind evaluation, where experts rate model explanations on clarity, relevance, and reasoning quality without knowing whether the explanation was human- or model-generated. This approach reduces bias and provides insight into how model explanations are perceived in practice.

V. Results and Discussion

The results demonstrate that LLMs are capable of producing coherent and contextually relevant qualitative explanations. In many cases, model-generated explanations closely resemble expert reasoning, particularly for well-defined themes and explicit sentiment cues. This suggests that LLMs can serve as effective assistants in exploratory qualitative analysis.

However, faithfulness analysis reveals notable limitations. While most explanations reference input content, a non-trivial proportion introduce generalized assumptions or inferred intentions not directly supported by the text. These hallucinated rationales pose risks in sensitive applications where traceability is critical.

Consistency evaluation shows that explanations can vary substantially across prompt rephrasings, even when final conclusions remain the same. This indicates that explanations are not always anchored to stable internal representations, raising concerns about their reliability as interpretive tools.

Human alignment scores indicate that experts generally find LLM explanations readable and logically structured. Nevertheless, experts frequently note a lack of depth in causal reasoning,

with models favoring surface-level interpretations over deeper contextual analysis [12]. This highlights the gap between linguistic fluency and true qualitative understanding.

Overall, the findings suggest that while LLMs are promising qualitative analysts, their explanations should be treated as hypotheses rather than authoritative interpretations. Human oversight remains essential, particularly in domains requiring nuanced judgment.

VI. Limitations and Ethical Considerations

One limitation of this study is the reliance on expert judgment, which itself is subjective. While this reflects the nature of qualitative analysis, it also complicates evaluation and generalization. Future work could explore cross-cultural or interdisciplinary expert panels to broaden interpretive perspectives.

Another limitation lies in the opacity of internal model representations. Because LLMs do not expose their internal reasoning mechanisms, explanations remain post-hoc narratives rather than verifiable causal accounts. This restricts the degree of trust that can be placed in model-generated justifications. Ethical concerns arise when LLM explanations are used to support decisions affecting individuals or communities. Persuasive but flawed explanations may reinforce biases or legitimize incorrect interpretations. Transparency about model limitations is therefore essential in deployment contexts.

There is also a risk of over-reliance on LLMs for qualitative research, potentially marginalizing human expertise. These systems should be positioned as augmentative tools rather than replacements for human analysts. Finally, data biases present in training corpora can subtly influence qualitative interpretations. Without careful monitoring, models may reproduce dominant narratives while overlooking minority perspectives.

VII. Conclusion

Large Language Models offer a powerful new paradigm for qualitative analysis, capable of generating explanations that are fluent, structured, and often aligned with human reasoning. However, explainability in this context extends beyond surface-level justification and requires careful evaluation of faithfulness, consistency, and interpretive depth. This study demonstrates that while LLMs can meaningfully support qualitative analysis, their

explanations are best viewed as interpretive aids rather than definitive reasoning processes. Advancing explainable qualitative analysis will require continued research into grounding mechanisms, stability of reasoning, and human-centered evaluation frameworks, ensuring that these models are deployed responsibly and transparently in real-world decision-making contexts.

REFERENCES:

[1] S. Khairnar and D. Bodra, "A Data-Driven Approach to Air Traffic Delay Prediction and Sentiment Evaluation," *International Journal of Basic and Applied Sciences,* vol. 14, no. 4, pp. 184-193, 2025.

[2] J. Yu *et al.*, "The Shadow of Fraud: The Emerging Danger of AI-powered Social Engineering and its Possible Cure," *arXiv preprint arXiv:2407.15912,* 2024.

[3] W. Guo, S. Zong, S. Chen, F. Zhao, and Y. Shang, "Design and Implementation of a New Serverless Conversational Survey System," in *2021 IEEE International Conference on Data Science and Computer Application (ICDSCA)*, 2021: IEEE, pp. 358-363.

[4] E. Aghaei, X. Niu, W. Shadid, and E. Al-Shaer, "Securebert: A domain-specific language model for cybersecurity," in *International Conference on Security and Privacy in Communication Systems*, 2022: Springer, pp. 39-56.

[5] S. Khairnar and D. Bodra, "Analysis and Evaluation of Modern Lightweight Cryptographic Algorithms: Standards, Hardware Implementation, and Security Considerations," *International Journal of Computer Applications,* vol. 975, p. 8887, 2025.

[6] F. Zhao and F. Yu, "Enhancing multi-class news classification through Bert-augmented prompt engineering in large language models: A novel approach," in *The 10th International scientific and practical conference "Problems and prospects of modern science and education"(March 12–15, 2024) Stockholm, Sweden. International Science Group. 2024. 381 p.*, 2024, p. 297.

[7] S. Al-Mansoori and M. B. Salem, "The role of artificial intelligence and machine learning in shaping the future of cybersecurity: trends, applications, and ethical considerations," *International Journal of Social Analytics,* vol. 8, no. 9, pp. 1-16, 2023.

[8] A. Nassar and M. Kamal, "Ethical dilemmas in AI-powered decision-making: a deep dive into big data-driven ethical considerations," *International Journal of Responsible Artificial Intelligence,* vol. 11, no. 8, pp. 1-11, 2021.

[9] F. Zhao, Y. Shang, and T. J. Trull, "FENAP: Foundation models for EMA-derived narrative analysis and prediction," *Journal of Healthcare Informatics Research,* pp. 1-21, 2024.

[10] F. Zhao, F. Yu, and Y. Shang, "A New Method Supporting Qualitative Data Analysis Through Prompt Generation for Inductive Coding," in *2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI)*, 2024: IEEE, pp. 164-169.

[11] L. Weidinger *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359,* 2021.

[12] F. Zhao, F. Yu, T. Trull, and Y. Shang, "A new method using LLMs for keypoints generation in qualitative data analysis," in *2023 IEEE Conference on Artificial Intelligence (CAI)*, 2023: IEEE, pp. 333-334.