A Literature Review of LLM-as-a-Judge Methods for Evaluating 5W1H Information Extraction in the Context of Requirements Engineering

José Cassola Bacallao[0000−0002−1029−1548], Ítalo Donoso Barraza[0009−0006−0468−9919], and Brian Keith Norambuena[0000−0001−5734−8962]

Abstract The evaluation of information extraction systems has traditionally relied on lexical similarity metrics such as ROUGE and BLEU, which are insufficient to capture semantic accuracy and contextual completeness. This literature review examines the emerging paradigm of LLM-as-a-Judge for evaluating natural language processing tasks and explores its potential intersection with 5W1H (Who, What, When, Where, Why, How) information extraction in requirements engineering contexts. The review was conducted using a structured exploratory approach, combining targeted searches across leading digital libraries with thematic organization of findings. We analyze recent advances in LLM-based evaluation methodologies, review current approaches to 5W1H extraction and evaluation, and identify gaps in applying these techniques to requirements engineering. Our review reveals that LLM-based evaluation methods achieve a significantly higher correlation with human judgment compared to traditional metrics, while providing explainable assessments that could benefit requirements engineering. We synthesize findings from multiple research streams to understand how LLM-as-a-Judge approaches might address the unique challenges of evaluating information extraction from formal specifications, considering stakeholder diversity and regulatory compliance needs. The literature reveals promising, but largely unexplored, opportunities at the intersection of these three domains. This work contributes to understanding how advances in artificial intelligence could improve software process improvement by highlighting current research gaps and potential synergies between LLM evaluation methods, structured information extraction, and requirements engineering practices.

José Cassola Bacallao
Universidad Católica del Norte, Av. Angamos 0610, e-mail: jose.cassola@alumnos.ucn.cl

Ítalo Donoso Barraza
Universidad Católica del Norte, Av. Angamos 0610, e-mail: italo.donoso@ucn.cl

Brian Keith Norambuena
Universidad Católica del Norte, Av. Angamos 0610, e-mail: brian.keith@ucn.cl, *Corresponding Author*

1 Introduction

Extracting and evaluating relevant content from requirements documents, user stories, and technical specifications is a complex problem [1]. Information extraction, particularly through the 5W1H framework (Who, What, When, Where, Why, How), has emerged as a fundamental approach to structuring unstructured text [2] and could be used to ensure the completeness of the requirements [3, 4]. However, the evaluation of such extraction systems remains problematic, as traditional metrics are insufficient to capture the semantic richness and contextual accuracy required in software engineering contexts.

The 5W1H framework, originally developed for journalism, provides a systematic approach to capture essential information elements [5]. In requirements engineering, this framework helps ensure that specifications address all critical aspects: stakeholders (Who), functionality (What), temporal constraints (When), system boundaries (Where), rationale (Why), and implementation approach (How) [6]. Despite its importance, current evaluation methods for 5W1H extraction rely primarily on lexical overlap metrics that inadequately assess semantic correctness and completeness.

The LLM-as-a-Judge paradigm [7] leverages the semantic understanding capabilities of LLMs to evaluate content through methods that align more closely with human judgment. LLM-based evaluation demonstrates correlation coefficients exceeding 0.5 with human assessments, substantially outperforming traditional metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which typically achieve correlations below 0.3 [8].

Publications were identified through systematic searches in the ACM Digital Library, IEEE Xplore, SpringerLink, and arXiv, covering the period 2018-2025. The search strategy combined keywords representing three thematic axes: *LLM-as-a-Judge*, *5W1H information extraction*, and *requirements engineering*. Studies were then organized thematically to highlight overlaps, differences, and gaps across domains. This method became appropriate for an exploratory review, where the goal is to identify research opportunities and conceptual synergies across emerging fields.

This literature review examines the intersection of three domains: **(1)** LLM-as-a-Judge evaluation methodologies, **(2)** 5W1H information extraction techniques and **(3)** requirements engineering practices. We analyze how LLM-based evaluation can address current limitations in assessing information extraction quality, with specific attention to the unique challenges of requirements documents. Our review contributes to the growing body of work on AI-driven software process improvement by analyzing the potential application of LLM-as-a-Judge methods to evaluate 5W1H extractors in requirements engineering contexts. Figure 1 shows the core contributions of our work.

The remainder of this paper is organized as follows: Section 2 provides a background on evaluation metrics and their limitations. Section 3 reviews the LLM-as-a-Judge paradigm and current frameworks. Section 4 examines the 5W1H extraction methods and evaluation approaches. Section 5 analyzes applications in requirements engineering. Section 6 provides an integrated discussion on LLM-based evaluation of 5W1H extractors for requirements engineering. Finally, Section 7 concludes with future research directions.

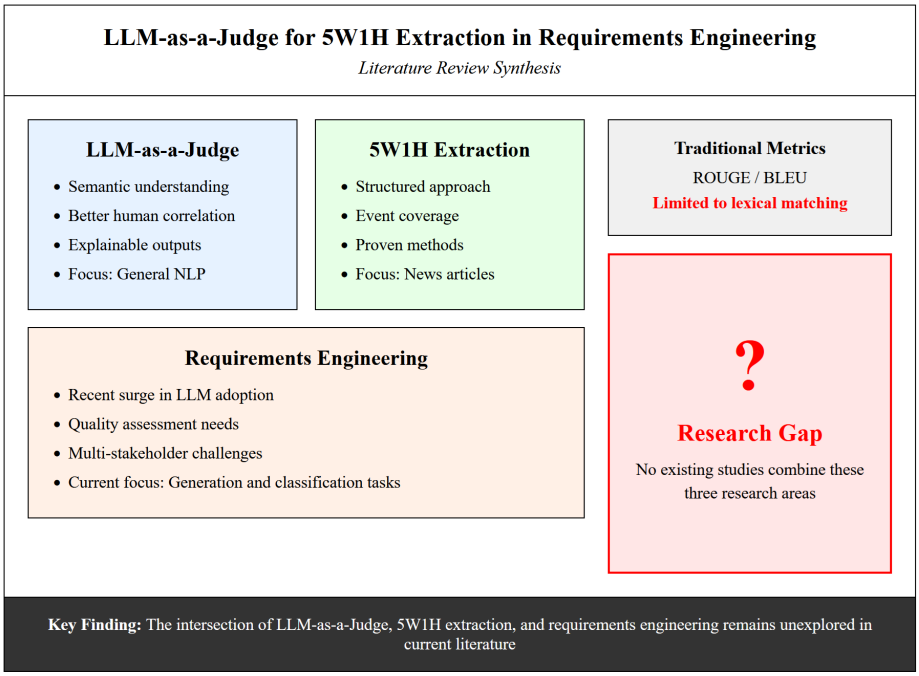

Fig. 1 Key highlights of the literature review of LLM-as-a-Judge for 5W1H extraction in the context of requirements engineering.

2 Traditional Evaluation Metrics and Their Limitations

BLEU (Bilingual Evaluation Understudy) is a metric introduced by Papineni et al. [9] that measures the *n*-gram overlap between the generated and reference texts. Originally designed for machine translation, BLEU has been widely adopted for various NLP tasks despite fundamental limitations. The metric operates on surface-level lexical matching, treating synonyms as entirely different words and ignoring semantic equivalence [10].

For 5W1H extraction evaluation, BLEU's limitations become particularly acute. Consider evaluating the extraction of "Who" information: BLEU would penalize "CEO" and "Chief Executive Officer" as different, despite their semantic equivalence. Song et al. [11] demonstrated that BLEU scores often show poor correlation with human judgments for tasks requiring semantic understanding, with correlation coefficients frequently below 0.3. This limitation becomes even more pronounced when evaluating complex questions such as "Why" and "How", where answers may be paraphrased or distributed across multiple sentences.

ROUGE encompasses multiple metrics that focus on the recall of *n*-grams, longest common subsequences, and skip-bigrams [12]. Although ROUGE-L's consideration of sentence-level structure represents an improvement over BLEU, it remains fundamentally limited to surface-level matching [13]. The metric calculates overlap between

candidate and reference texts without understanding whether differences represent genuine semantic distinctions or merely stylistic variations.

Akter et al. [14] conducted a comprehensive analysis of ROUGE's performance on extractive summarization tasks, finding that the metric fails to distinguish between semantically equivalent paraphrases and genuinely different content. For requirements engineering contexts, where technical terminology may vary across stakeholders, this limitation proves particularly problematic. A requirement stating "*The system shall authenticate users*" would score poorly against "*User authentication must be performed by the system*" despite expressing identical functionality.

METEOR (Metric for Evaluation of Translation with Explicit ORdering) attempted to address some limitations by incorporating stemming, synonymy, and paraphrase matching [15]. The metric combines precision and recall with a penalty for fragmentation, aiming to reward fluent translations that preserve meaning. Lavie and Denkowski [16] showed that METEOR achieves better correlation with human judgment than BLEU, typically reaching 0.4-0.5 correlation coefficients. However, METEOR's reliance on external resources like WordNet limits its applicability across domains and languages. Saadany and Orasan [17] found that METEOR's performance degrades significantly when applied to specialized domains with technical vocabulary not well-represented in general-purpose synonym databases. For requirements documents containing domain-specific terminology, this limitation significantly impacts evaluation accuracy. Terms like "shall" and "must" in requirements contexts carry specific regulatory meanings that general-purpose synonym databases fail to capture.

The fundamental issue underlying all traditional metrics is their focus on lexical rather than semantic similarity. Barnes et al. [18] recently demonstrated this limitation in their evaluation of Spanish and Basque summarization systems, where traditional metrics showed correlation coefficients below 0.3 with human judgments for semantic criteria like information completeness and coherence. Their work particularly highlighted how the 5W1H criterion, when evaluated by traditional metrics, failed to capture whether extracted information actually answered the intended questions.

For the evaluation of the extraction of 5W1H, the semantic gap manifests itself in several critical ways. Paraphrase blindness means that semantically equivalent extractions receive low scores due to lexical differences, undermining the evaluation of systems that correctly identify information but express it differently. Context ignorance prevents metrics from assessing whether extracted information is contextually appropriate for the specific W or H question that is being answered. Completeness insensitivity allows missing critical information to go undetected if other elements match well lexically. Coherence neglect occurs when extracted elements score well individually, yet form an incoherent or contradictory whole when combined. Such limitations underscore the need for evaluation methods capable of semantic understanding beyond surface-level matching. **LLM-based evaluation methods** address this gap by incorporating a deeper semantic analysis into assessment tasks [19].

3 The LLM-as-a-Judge Paradigm

The LLM-as-a-Judge paradigm leverages large language models' semantic understanding to assess text quality [7]. LLM judges evaluate semantic accuracy, contextual appropriateness, and qualitative aspects of text—capabilities that emerge from extensive pre-training on diverse corpora, which develops their implicit understanding of language quality, factual precision, and coherence [20].

LLMs trained on human-generated text internalize human preferences and quality standards. When properly prompted, these models can serve as proxies for human judgment, offering consistent and scalable evaluation capabilities. Gu et al. [19] provide a comprehensive taxonomy of LLM-as-a-Judge applications, identifying key deployment patterns across evaluation tasks.

Liu et al. [8] introduced G-Eval, employing *Chain-of-Thought* (CoT) prompting to enhance LLM evaluation capabilities. The framework establishes clear task definitions with evaluation criteria tailored to the assessment context, generates evaluation steps through CoT prompting to break down complex assessments, and produces scores using probability-weighted token likelihoods.

G-Eval achieves Spearman correlation coefficients of 0.514 with human judgments on summarization tasks, significantly outperforming ROUGE-L (0.289) and BERTScore (0.392) [8]. By decomposing complex evaluation tasks into manageable steps, the framework mirrors human evaluation processes. For 5W1H extraction evaluation, G-Eval's CoT process can explicitly verify each dimension and overall coherence.

Chan et al. [21] proposed ChatEval, which employs multiple LLM agents with diverse perspectives to evaluate text through structured debate. The system instantiates agents with distinct personas—domain experts, end users, technical reviewers—who engage in structured debates following predefined protocols. Consensus-building mechanisms aggregate individual assessments into final scores.

ChatEval demonstrates superior performance on subjective quality assessments, particularly where multiple valid interpretations exist. The multi-agent architecture proves valuable when evaluating requirements documents, where stakeholders prioritize different quality aspects. A developer agent might focus on technical feasibility while a user experience agent emphasizes clarity and completeness.

Recent work has developed purpose-built evaluation models rather than repurposing general LLMs. JudgeLM [22] fine-tunes smaller models (7B to 33B parameters) specifically for evaluation tasks. Training on 100K judge samples with high-quality GPT-4 judgments creates specialized models that match or exceed general-purpose LLM performance while requiring fewer computational resources.

Prometheus 2 [23] provides an open-source alternative based on Llama-2-Chat. While requiring reference answers and score rubrics, Prometheus achieves evaluation performance comparable to GPT-4 at reduced cost. The model's training incorporates diverse evaluation scenarios and explicit rubrics, enabling generalization across assessment contexts. These developments democratize access to high-quality evaluation capabilities, particularly for academic research and resource-limited organizations working on information extraction systems.

LLM-as-a-Judge faces several documented challenges. Huang et al. [24] conducted extensive experiments revealing systematic biases in LLM judges: preferences for longer responses, sensitivity to prompt variations, and tendencies to favor certain writing styles. Their work demonstrates that fine-tuned judge models often function as task-specific classifiers rather than general evaluators, potentially limiting cross-domain transferability.

Chen et al. [25] documented additional biases through controlled experiments. Position bias causes responses appearing first to receive higher scores. Authority bias favors responses citing sources or using technical language. Beauty bias prefers well-formatted text regardless of content quality. These biases mirror human evaluation tendencies but may be amplified in automated systems.

Computational cost poses practical constraints. LLM-based evaluation requires substantial computational resources compared to traditional metrics, potentially limiting real-time applications. This cost-performance tradeoff becomes particularly relevant for continuous integration environments where requirements undergo frequent evaluation during development.

The explainability paradox presents additional complexity. While LLMs generate detailed explanations for their judgments, Leiter et al. [26] found that these explanations do not always faithfully represent the actual decision process. This disconnect between generated rationales and underlying reasoning mechanisms complicates efforts to debug or improve evaluation systems.

4 5W1H Information Extraction: Methods and Evaluation

The 5W1H framework, originally a manual journalistic practice, has been successfully automated in recent extraction systems. Hamborg et al. [2] developed Giveme5W1H, achieving 73% overall precision and 82% for the four W questions (Who, What, When, Where). Their system employs carefully crafted syntactic and semantic rules, leveraging dependency parsing and named entity recognition to identify answer candidates. However, the rule-based approach struggles with complex questions requiring inference, particularly Why and How, which often demand understanding of causal relationships and procedural knowledge not explicitly stated in the text.

Keith et al. [6] advanced the field by leveraging 5W1H extraction to evaluate inverted pyramid structure in news articles, introducing the Inverted Pyramid Score (IPS). Their work demonstrated that 5W1H elements serve as indicators of information organization and completeness, with applications beyond simple extraction. By analyzing the location and quality of 5W1H answers within articles, they could assess adherence to journalistic standards and information structuring principles.

Recent advances by Cao et al. [27] and Muñoz et al. [28] represent a paradigm shift toward LLM-based extraction. For example, the study of Cao et al. created a dataset of 3,500 manually annotated entries across four news corpora, revealing the superiority of fine-tuned models over general-purpose systems. Their experiments with LLaMA, Vicuna, and Guanaco models using QLoRA (Quantized Low-Rank Adaptation) demonstrate that specialized models outperform general-purpose ChatGPT, particularly for

contextually complex questions. The performance gap is most pronounced for Why and How questions, where fine-tuned models achieve 15-20% higher accuracy through better understanding of implicit causal relationships and procedural descriptions.

Traditional evaluation of 5W1H extraction relies on precision, recall, and F1-score calculated through exact or partial matching against gold-standard annotations. Hamborg et al. [29] introduced a more nuanced evaluation framework that goes beyond binary correct/incorrect classifications. Their system categorizes extractions as correct when they completely match the gold standard, partial when they overlap but incompletely capture the information, missing when gold standard elements are not extracted, incorrect when extracted elements are not present in the gold standard, and spurious when extracted elements are irrelevant to the specific W or H question.

This categorical approach provides more insight than simple accuracy metrics but still emphasizes surface matching over semantic accuracy. The challenge becomes particularly pronounced for Why and How questions, where answers may be distributed across multiple sentences or require inference from implicit information. A causal explanation might be correctly identified but expressed differently from the gold standard, resulting in low scores despite semantic correctness.

Another approach is the shift toward semantic evaluation, which represents a fundamental change in how we assess extraction quality. BERTScore [30] pioneered this approach by computing similarity using contextual embeddings rather than surface-level matching. The metric leverages pre-trained BERT models to capture semantic equivalence between paraphrases and related concepts, achieving significantly better correlation with human judgments than traditional metrics. However, BERTScore still operates primarily at the sentence level without considering document-wide coherence or completeness. For 5W1H extraction, this means the metric might correctly identify that individual elements are semantically similar to references but miss that crucial information is absent or that extracted elements form an incomplete picture.

5 Applications in Requirements Engineering

The integration of LLMs into requirements engineering represents one of the most rapid adoptions of AI technology in software engineering. A systematic literature review by Ahmad et al. [43] identified 35 research papers published between 2023-2024 focusing specifically on LLM applications in requirements engineering. This explosive growth reflects both the potential of LLMs to address longstanding challenges and the pressing need for improved requirements processing tools in complex software systems.

The applications span the entire requirements engineering lifecycle. In requirements elicitation, LLMs could assist in extracting requirements from stakeholder interviews and informal descriptions [31]. During analysis, they automatically categorized functional versus non-functional requirements with accuracy exceeding 90% in controlled studies [32]. In general, LLMs could be used to analyze requirement smells [33] or to aid in requirements management to assess the impact across related specification when changes are introduced [34].

ISO/IEC 29148 defines quality characteristics for requirements including completeness, consistency, correctness, and clarity [35]. Traditional approaches to assessing these characteristics rely on manual inspection or rule-based tools with limited semantic understanding. Recent work by Lubos et al. [36] demonstrates that LLMs can assess these characteristics with accuracy comparable to human experts while providing detailed explanations for their assessments. These capabilities extend beyond simple rule checking to understand the semantic content and implications of requirements. For instance, an LLM could identify that two requirements conflict not because they use contradictory keywords but because their logical implications are incompatible when considered in the system context.

Technical terminology presents another challenge, as requirements documents incorporate domain-specific vocabulary that may not appear in general-purpose training corpora [37]. A requirement for an avionics system might reference "DO-178C compliance" or "ARINC 429 bus protocols," terms that require specialized knowledge to evaluate properly. This challenge is compounded when requirements span multiple technical domains, requiring evaluation methods to adapt to varied vocabularies. Furthermore, stakeholder diversity adds complexity to requirements evaluation [38]. Different stakeholders interpret requirements through their own perspectives and priorities. A requirement that seems clear to a developer might be ambiguous to a tester or end-user. Evaluation methods must account for these multiple viewpoints rather than assuming a single correct interpretation.

Regulatory compliance requirements add another dimension to evaluation complexity [39]. Requirements must often satisfy specific regulatory standards, with precise wording required for compliance. Evaluation methods must understand not just whether requirements are clear and complete but whether they meet regulatory obligations.

While 5W1H extraction has been extensively studied for news articles, its application to requirements engineering remains surprisingly underexplored [3, 4]. Yet requirements documents contain all the information elements of the 5W1H framework, albeit distributed and expressed differently than in journalistic text. The **Who** element appears as stakeholders, actors, and responsible parties throughout requirements documents. The **What** manifests as functions, features, and capabilities the system must provide. **When** is expressed through deadlines, milestones, and temporal constraints on system behavior. **Where** encompasses system boundaries, deployment contexts, and integration points with other systems. **Why** provides business rationale, goals, and justifications for specific requirements. **How** describes implementation constraints, quality attributes, and technical approaches.

However, this information is often distributed across documents rather than concentrated in specific sections. Technical specifications might separate functional requirements from quality attributes, even though both contribute to understanding **What** the system must do and **How** it must perform. Requirements traceability matrices link requirements to their rationales (**Why**) but in tabular rather than narrative form. Use cases describe **Who** interacts with the system and **What** they accomplish but may scatter **When** and **Where** constraints throughout scenario descriptions.

Thus, it would be expected that traditional extraction methods designed for narrative text struggle with these characteristics. Rule-based systems that expect concentrated information fail when faced with distributed specifications. Statistical methods trained

in news corpora cannot adapt to the formal language and structure of requirements documents. These limitations suggest the need for extraction methods that understand the unique characteristics of requirements documentation while maintaining the comprehensive coverage provided by the 5W1H framework.

Beyond the technical challenges of applying 5W1H extraction techniques to requirements documents, it is essential to consider an additional dimension: the social and human factors inherent to the requirements engineering process. Elements such as communication ambiguity, tacit knowledge, linguistic and cultural barriers, resistance to change, and emotional or motivational aspects significantly influence the quality of the obtained requirements [40, 41, 42]. For example, when stakeholders employ domain-specific technical jargon, provide imprecise descriptions due to organizational tensions, or interact in distributed contexts with cultural differences, requirements may appear complete on the surface but lack the necessary depth for precise 5W1H extraction.

Likewise, the definition of a single "correct" set of answers for the 5W1H becomes questionable in multi-stakeholder scenarios. The same requirement can be interpreted differently depending on the role: while a developer may focus on technical aspects (How), the quality assurance team prioritizes testability (What), and end users primarily value business impact (Why). These interpretations, influenced by diverse contexts, priorities, and cultural frameworks, should not be considered mutually exclusive. On the contrary, LLM-based evaluators should recognize the validity of multiple simultaneous readings, understanding that a comprehensive understanding of requirements emerges from the integration of these perspectives. Furthermore, factors such as team motivation [42] directly affect the quality and completeness of the extracted information.

6 Discussion

Cross-Domain Applications and Gaps. The literature on LLM-as-a-Judge has mainly focused on general NLP tasks such as summarization, translation, and evaluation of dialogues [8, 7]. Meanwhile, 5W1H extraction research has focused almost exclusively on news articles [2, 27], with limited exploration in other domains. Requirements engineering studies have begun adopting LLMs for various tasks [43], but the intersection of these three areas remains largely unexplored.

Barnes et al. [18] represent one of the few studies explicitly using 5W1H as an evaluation criterion, although their focus on summarization differs substantially from information extraction. Their finding that GPT-4o achieved a strong correlation with human judgment for 5W1H completeness suggests potential for broader applications, yet they do not explore requirements engineering contexts.

The gap becomes more apparent when the requirements engineering literature is examined. Although studies like those of Lubos et al. [36] demonstrate that LLMs evaluate the quality of requirements, they focus on traditional quality attributes rather than structured information extraction. Similarly, Ferrari et al. [34] identify information patterns in requirements but do not connect to modern LLM evaluation approaches.

Domain-Specific Challenges. Multiple researchers have independently identified the challenges that would affect any attempt to apply LLM-based evaluation in the

context of requirements. Berry [37] extensively documents how requirements documents differ from general text in vocabulary, structure, and implicit assumptions. These differences suggest that direct application of 5W1H extractors trained in news would likely fail. The vocabulary challenge appears consistently in all studies. Saadany and Orasan [17] found that even METEOR, designed to handle synonyms, struggles with domain-specific terminology. Requirements documents compound this problem with their formal language, where terms such as "shall" carry precise regulatory meanings absent from the general corpora [39]. Structural differences present another recurring theme. Pohl [44] describes how requirement documents use hierarchical numbering, extensive cross-references, and tabular formats that differ fundamentally from narrative text. This contrasts sharply with the inverted pyramid structure of classical news [6].

Evaluation Complexity in Multi-Stakeholder Contexts. The requirements engineering literature consistently emphasizes stakeholder diversity as a defining characteristic [38]. This creates evaluation challenges not addressed in current LLM-as-a-Judge research, which typically assumes a single quality standard. ChatEval [21] introduces multi-agent evaluation, but their agents represent different aspects of quality rather than genuinely different stakeholder perspectives. Research on requirements quality reveals how different stakeholders interpret the same requirement differently. Femmer et al. [33] found that developers and testers often disagree on requirement clarity, not due to evaluation error but because they need different information. This suggests that any evaluation approach must accommodate multiple valid interpretations rather than seeking a single correct assessment. The implications extend to 5W1H extraction evaluation. What constitutes a complete "Who" extraction might vary between stakeholders—developers need technical roles while business analysts focus on business stakeholders. Current evaluation approaches, whether traditional metrics or LLM-based, do not account for this fundamental multiplicity of perspectives.

Technical and Methodological Considerations. Recent studies reveal both opportunities and challenges in applying LLMs to specialized domains. Huang et al. [24] demonstrate that fine-tuned LLM judges often become task-specific classifiers, losing generalization ability. This finding suggests that creating requirements-specific evaluators might sacrifice the flexibility that makes LLMs attractive. Computational cost emerges as a practical concern across multiple studies. While Kim et al. [23] show that smaller models can match GPT-4 performance when properly trained, the training process itself requires substantial resources. For organizations with limited budgets, this creates barriers to adopting advanced evaluation methods. The explainability paradox identified by Leiter et al. [26] has particular relevance for requirements engineering. They found that LLM-generated explanations do not always reflect actual decision processes. In regulated industries where requirements traceability is mandatory, this disconnect between explanation and reasoning could create compliance risks.

Insights from Related Applications. The LACA framework [45] demonstrates how LLMs can support qualitative analysis when combined with human expertise. Their approach of using LLMs for initial coding followed by human validation might address some reliability concerns in requirements evaluation. However, they focus on thematic analysis rather than structured information extraction. Research on requirements classification [32] shows that even well-established categories like functional versus non-functional requirements challenge LLMs without specific training. This suggests

that the more complex task of evaluating 5W1H extraction quality would face similar or greater challenges.

Potential Impact on Software Quality. Improved evaluation of 5W1H extraction directly impacts software quality through several interconnected mechanisms. Early defect detection represents perhaps the most significant benefit, as identifying incomplete or ambiguous requirements before they propagate to design and implementation phases prevents costly rework. Studies by Boehm and Basili [46] demonstrate that fixing requirements defects costs 10-100 times less than fixing the same defects discovered during testing or production. LLM-based evaluation naturally aligns with established software engineering standards and maturity models, facilitating adoption and demonstrating value within existing quality frameworks. The multi-dimensional evaluation approach enabled by LLM-as-a-judge models maps directly to software quality characteristics defined in ISO/IEC 25010 [47].

Limitations. Several limitations merit careful consideration when interpreting our findings. Computational costs remain a significant consideration for practical deployment of LLM-as-a-judge models for 5W1H extraction in requirements engineering. Although model efficiency continues to improve, LLM-based evaluation still requires substantially more computational resources than traditional metrics. Bias and consistency in LLM judges represent ongoing challenges requiring continued research. Position bias, length preference, and other systematic biases identified in recent studies must be addressed through a careful prompt design and possibly ensemble methods.

7 Conclusions and Future Research Directions

This literature review has examined the convergence of three important areas: **(1)** LLM-as-a-Judge evaluation methodologies, **(2)** 5W1H information extraction, and **(3)** requirements engineering practices. Thus, this review contributes to the intersection of AI and software process improvement. We provide the first analysis connecting LLM-as-a-Judge methodologies with 5W1H extraction evaluation in requirements engineering contexts, identifying synergies and adaptation requirements.

First, our review identified a significant gap in applying 5W1H extraction and evaluation to requirement engineering contexts. Although the journalism domain has seen substantial advancement in automated 5W1H extraction, requirements engineering has not fully leveraged these capabilities despite clear applicability. Requirements documents contain all 5W1H elements, but distribute them differently than news articles, necessitating specialized approaches for both extraction and evaluation. Second, the rapid embrace of LLM technologies by the requirements engineering community, demonstrates readiness for advanced AI-assisted methods. However, most of the work focuses on the generation or classification tasks rather than the extraction and evaluation, indicating an opportunity for a significant contribution.

In terms of future work, this review identifies several promising avenues for future research. Empirical validation is the most immediate need, which requires implementation of LLM-as-a-judge models for 5W1H extraction techniques and systematic comparison against traditional metrics and human judgment across diverse requirements datasets.

Such studies should examine not only the correlation with human assessment, but also practical impacts on the quality and development outcomes of requirements.

Additionally, a particularly promising avenue for future research lies in training LLMs to detect and evaluate social and human factors within requirements. This includes developing specialized metrics that assess not only technical extraction accuracy but also the presence of social and human barriers such as linguistic ambiguity, cultural differences, and motivational factors. Such capabilities would be especially valuable for agile approaches.

Human-AI collaboration research must identify optimal combinations of automated evaluation and human review. Key questions include when human intervention provides the most value, how to present evaluation results for efficient human review, and how to incorporate human feedback to improve automated evaluation. The goal is augmenting rather than replacing human judgment. The future of requirements engineering lies not in replacing human judgment, but in augmenting it with AI capabilities that handle routine evaluation tasks while flagging complex issues for human attention. This human-AI partnership promises to elevate requirements engineering from a necessary but often painful process to a value-adding activity that measurably improves software outcomes.

Acknowledgment

This research is funded by the ANID FONDECYT 11250039 Project. The author is also supported by Project 202311010033-VRIDT-UCN.

References

1. Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.
2. Hamborg, F., Breitinger, C., & Gipp, B. (2019). Giveme5W1H: A universal system for extracting main events from news articles. arXiv preprint arXiv:1909.02766.
3. Jabar, M. A., Ahmadi, R., Shafazand, M. Y., Ghani, A. A. A., Sidi, F., & Hasan, S. A. (2013, August). An automated method for requirement determination and structuring based on 5W1H elements. In 2013 IEEE 4th Control and System Graduate Research Colloquium.
4. Pabuccu, Y. U., Yel, I., Helvacioglu, A. B., & Asa, B. N. (2022). The requirement cube: A requirement template for business, user, and functional requirements with 5w1h approach. International Journal of Information System Modeling and Design (IJISMD), 13(1), 1-18.
5. Harrower, T. (2010). Inside reporting. McGraw-Hill Education.
6. Keith Norambuena, B., Horning, M., & Mitra, T. (2020). Evaluating the inverted pyramid structure through automatic 5W1H extraction and summarization. In Proceedings of Computation + Journalism Symposium.
7. Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. Advances in Neural Information Processing Systems, 36, 1-29.
8. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 2511-2522).

9. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).
10. Culy, C., & Riehemann, S. Z. (2003). The limits of n-gram translation evaluation metrics. In Proceedings of Machine Translation Summit IX: Papers.
11. Song, X., Cohn, T., & Specia, L. (2013). BLEU deconstructed: Designing a better MT evaluation metric. International Journal of Computational Linguistics and Applications, 4(2), 29-44.
12. Lin, C. Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
13. Barbella, M., & Tortora, G. (2022). Rouge metric evaluation for text summarization techniques. SSRN Electronic Journal.
14. Akter, M., Bansal, N., & Karmaker, S. K. (2022). Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE?. In ACL 2022 (pp. 1547-1560).
15. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).
16. Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. Machine translation, 23(2), 105-115.
17. Saadany, H., & Orăsan, C. (2021). BLEU, METEOR, BERTScore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. In Proceedings of the Translation and Interpreting Technology Online Conference (pp. 48-56).
18. Barnes, J., Perez, N., Bonet-Jover, A., & Altuna, B. (2025). Summarization metrics for Spanish and Basque: Do automatic scores and LLM-judges correlate with humans?. arXiv preprint arXiv:2503.17039.
19. Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... & Guo, J. (2024). A survey on LLM-as-a-Judge. arXiv preprint arXiv:2411.15594.
20. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 15(3), 1-45.
21. Chan, C. M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., ... & Liu, Z. (2023). ChatEval: Towards better LLM-based evaluators through multi-agent debate. arXiv preprint arXiv:2308.07201.
22. Zhu, L., Wang, X., & Wang, X. (2023). JudgeLM: Fine-tuned large language models are scalable judges. arXiv preprint arXiv:2310.17631.
23. Kim, S., Suk, J., Longpre, S., Lin, B. Y., Shin, J., Welleck, S., ... & Seo, M. (2024). Prometheus 2: An open source language model specialized in evaluating other language models. arXiv preprint arXiv:2405.01535.
24. Huang, H., Qu, Y., Liu, J., Yang, M., & Zhao, T. (2024). An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge models are task-specific classifiers. arXiv preprint arXiv:2403.02839.
25. Chen, G. H., Chen, S., Liu, Z., Jiang, F., & Wang, B. (2024). Humans or LLMs as the judge? A study on judgement biases. arXiv preprint arXiv:2402.10669.
26. Leiter, C., Opitz, J., Deutsch, D., Gao, Y., Dror, R., & Eger, S. (2023). The Eval4NLP 2023 shared task on prompting large language models as explainable metrics. In Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems (pp. 117-138).
27. Cao, Y., Lan, Y., Zhai, F., & Li, P. (2024). 5W1H extraction with large language models. arXiv preprint arXiv:2405.16150.
28. Muñoz, C., Mendoza, M., Lobel, H., & Keith, B. (2025, May). Imitating Human Reasoning to Extract 5W1H in News. In Companion Proceedings of ACM WWW 2025 (pp. 1199-1203).
29. Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., & Gipp, B. (2018). Giveme5W: Main event retrieval from news articles by extraction of the five journalistic W questions. In International Conference on Information (pp. 356-366).
30. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.
31. Ren, S., Nakagawa, H., & Tsuchiya, T. (2024, July). Combining prompts with examples to enhance llm-based requirement elicitation. In 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 1376-1381). IEEE.

32. Luitel, D., Hassani, S., & Sabetzadeh, M. (2024). Improving requirements completeness: Automated assistance through large language models. Requirements Engineering, 29(1), 73-95.
33. Femmer, H., Fernández, D. M., Wagner, S., & Eder, S. (2017). Rapid quality assurance with requirements smells. Journal of Systems and Software, 123, 190-213.
34. Ferrari, A., Gori, G., Rosadini, B., Trotta, I., Bacherini, S., Fantechi, A., & Gnesi, S. (2018). Detecting requirements defects with NLP patterns: an industrial experience in the railway domain. Empirical Software Engineering, 23(6), 3684-3733.
35. ISO/IEC/IEEE. (2018). ISO/IEC/IEEE 29148:2018 Systems and software engineering—Life cycle processes—Requirements engineering.
36. Lubos, F., Fischbach, J., Spies, M., & Vogelsang, A. (2024). Automated quality assessment of requirements using large language models. In 2024 IEEE 32nd International Requirements Engineering Conference (RE) (pp. 234-245).
37. Berry, D. M. (2017). Evaluation of tools for hairy requirements and software engineering tasks. In 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW), Lisbon, Portugal, 2017, pp. 284-291.
38. Glinz, M. (2007). On non-functional requirements. In 15th IEEE International Requirements Engineering Conference (RE 2007) (pp. 21-26).
39. Breaux, T. D., & Antón, A. I. (2008). Analyzing regulatory rules for privacy and security requirements. IEEE transactions on software engineering, 34(1), 5-20.
40. Donoso Barraza, Í., & Vega Zepeda, V. (2017). Factores sociales y humanos que afectan el proceso de educción de requerimientos: una revisión sistemática. RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação, 24, 69-83.
41. Branca, F., Matoff, P. F., Gaona, G., & Pérez, C. A. (2023). Incidencia de los factores humanos y socioculturales en la captura de requerimientos: una revisión de la literatura. In ASSE, Simposio Argentino de Ingeniería de Software.
42. Hidellaarachchi, D., Grundy, J., Hoda, R., & Mueller, I. (2023). The Influence of Human Aspects on Requirements Engineering-related Activities: Software Practitioners' Perspective. ACM Transactions on Software Engineering and Methodology, 32(5), 1-37.
43. Ahmad, A., Waseem, M., Liang, P., Fahmideh, M., Aktar, M. S., & Mikkonen, T. (2023). Towards human-bot collaborative software architecting with ChatGPT. In Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering (pp. 279-285).
44. Pohl, K. (2010). Requirements engineering: fundamentals, principles, and techniques. Springer Publishing Company.
45. Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P. Y. (2023). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In Companion Proceedings of IUI2023 (pp. 75-78).
46. Boehm, B., & Basili, V. R. (2001). Software defect reduction top 10 list. Computer, 34(1), 135-137.
47. ISO/IEC. (2011). ISO/IEC 25010:2011 Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models.