

Do LLMs Understand Validity? An Empirical Study of Machine-Generated Threats to Validity in Software Engineering

Ryan Dang ✉ 

Department of Informatics, University of California, Irvine, USA

Noshin Tahsin ✉ 

Department of Informatics, University of California, Irvine, USA

Thomas Zimmermann ✉ 

Department of Informatics, University of California, Irvine, USA

Abstract

Background. Threats to validity (TTV) are a core component of empirical research, yet software engineering studies fail to address them adequately. Published studies often omit validity discussions entirely, report generic concerns disconnected from the study’s actual methodology, or treat validity analysis as a procedural formality rather than a purposeful part of the research process.

Aim. This study investigates whether large language models (LLMs) can generate threats to validity that are meaningful, study-specific, and novel, i.e., not reported by the original authors.

Method. We conducted an empirical experiment using Gemini 2.5 Flash across 375 papers from ICSE 2025, withholding each paper’s TTV section prior to model input, ensuring the model reasoned from the study’s content rather than reproducing author-reported concerns. Generated threats were evaluated against a four-dimension rubric covering *relevance*, *specificity*, *clarity*, and *mitigation quality*, and compared against the original threats to determine *novelty*. All threats were then grouped into a taxonomy of seven categories to characterize the distribution of generated validity concerns.

Results. The experiment produced 2,673 threats across the dataset, of which 97% were rated as relevant, specific, and clearly articulated, and 92% received high ratings for mitigation quality. Furthermore, 61% of the generated threats were absent from the original papers, suggesting that the model identified validity concerns beyond those reported by the original authors.

Conclusions. LLMs can generate technically sound and previously unreported threats to validity at scale, demonstrating their capacity to support researchers in identifying overlooked methodological risks and strengthening the completeness of validity discussions, while augmenting rather than replacing researcher judgment in the process.

2012 ACM Subject Classification Software and its engineering; Computing methodologies → Natural language generation

Keywords and phrases Threats to Validity, Empirical Research, Large Language Models, LLMs, Software Engineering Research, Automated Analysis

Digital Object Identifier 10.4230/LIPIcs.ESEM.2026.76

Category Emerging Results, Vision & Reflection Track Paper

Supplementary Material *Software:* <https://zenodo.org/records/20453824>

1 Introduction

Empirical research is a significant area in software engineering, which offers a systematic basis for evaluating tools, methods, and practices in real-world contexts. The flexibility of empirical study design allows researchers to adapt their methodologies to their research



© Ryan Dang, Noshin Tahsin, and Thomas Zimmermann;
licensed under Creative Commons License CC-BY 4.0

20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026).

Editors: Robert Feldt, Maria Paasivaara, Daniel Mendez, Stefan Wagner, and Marvin Muñoz Barón;

Article No. 76; pp. 76:1–76:14



Leibniz International Proceedings in Informatics

LIPIcs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

questions. However, this flexibility can lead to the validity of findings not being uniform across studies. The threats to validity (TTV) section in research papers exist to address this, providing documented concerns about what could have influenced a study's results and how those risks were mitigated [7]. The threats serve as the primary mechanism by which authors reason about the reliability of their conclusions across dimensions like internal, external, construct, and conclusion validity [24, 22, 30, 14].

Despite their importance, TTV sections are often poorly handled in practice. Prior studies show that validity discussions can be shallow or absent, rely on standardized categories that may not fit the study method, and treat validity reasoning as an afterthought rather than integrating it into research design [22, 9, 25]. So far, reporting guidelines have been proposed to address this gap, guiding researchers in identifying where validity reasoning is missing [23, 28, 2, 20]. However, inconsistencies in TTV identification and reporting remain. This motivates us to investigate whether LLMs can provide additional support for identifying study-specific validity concerns. We aim to move beyond manual threat identification toward the automated identification and generation of study-specific, meaningful threats using LLMs given their growing use for reasoning and scientific writing tasks [4, 13, 1, 18]. To empirically investigate this potential, we aim to answer the following research questions (RQs).

- **RQ1:** How effectively can LLMs generate study-specific, meaningful threats to validity?
- **RQ2:** To what extent can LLMs identify threats that are not documented in the original paper?
- **RQ3:** What taxonomy of threats emerges from LLM-generated output and how well are different categories of threats represented?

To answer these, we conducted an empirical experiment using Gemini 2.5 Flash on 375 papers collected from ICSE 2025. For each paper, we removed the original TTV section and provided the remaining content to the model as input, asking it to generate threats based on the study content. This ensured that the model derived threats from the study's goals, methodology, and findings rather than reproducing author-reported concerns. The resulting 2,673 threats were evaluated using a rubric covering relevance, specificity, clarity, and mitigation quality, and compared with the original threats to assess novelty. Followed by that, all threats were grouped into a taxonomy of seven thematic categories to characterize the distribution of LLM-generated validity concerns.

The results show that 97% of the generated threats were rated as relevant, specific and clearly articulated, and 92% received high ratings for mitigation quality. Moreover, 61% of generated threats were absent from the original papers, demonstrating that the model was able to identify concerns the authors had not reported. These findings position LLMs as practical assistive tools for validity reasoning, contributing both an empirical characterization of LLM capability in this domain and a reusable taxonomy of validity concerns for software engineering research.

2 Related Work

Threats to validity are a foundational concern in empirical research and have been studied extensively across software engineering literature. Classic methodology texts establish the need to distinguish internal, external, construct, and conclusion validity, and to reason about them throughout study design, execution, and reporting [24, 22, 30, 14]. In software engineering, these principles have been adapted to different empirical research methods, with validity considerations varying according to the research design, goals, data sources, and

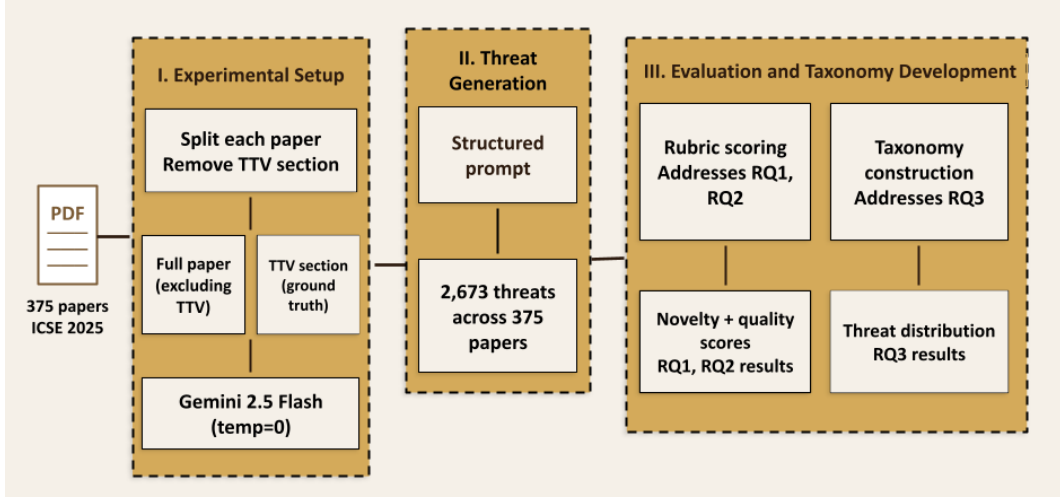
context [20, 29, 2]. Despite the availability of these methodological frameworks, prior research has identified persistent problems in how threats to validity are addressed and reported in practice [9, 25].

For example, a literature review of 405 papers from major software engineering venues (ICSE, ESEC/FSE, EMSE) found that 46% of empirical studies did not discuss threats to validity at all, and just 23% differentiated between types of validity [22]. Another study that analyzed 91 Distinguished ICSE Papers published from 2014 to 2023 found that only 22% of ACM papers discussed study-specific threats in depth and another 22% omitted them entirely [9]. Besides, TTV sections in research papers are often superficial, treated as a checklist or afterthought, rather than as an essential part of research design and interpretation [9, 25]. Studies also report misuse of validity categorization, issues with community practice, confusion with limitations and neglected tradeoffs between threats [25]. Findings show that researchers often use a standardized TTV categorization approach without considering if it fits their study type or method. Both authors and reviewers enforce shallow, standardized TTV discussions, sometimes requiring irrelevant threats to be discussed. Authors also frequently conflate study limitations with genuine validity threats. Additionally, there is little discussion of trade-offs between different types of validity threats, even though some may be more important than others depending on the study's goals. To improve the identification of TTV in practice, a range of structured taxonomies and guidelines have been proposed across the literature. For example, there are case-study reporting guidelines that explicitly call for threat identification across construct, internal, external, and conclusion validity [20]. However, despite this guidance, inconsistency in how threats are classified and reported has persisted over time with no observed trend of improvement [9].

Followed by work on guidelines for authors, studies have taken a step ahead to investigate whether the empirical research process can be enhanced through automation. Existing studies explored LLM's ability to infer, deduce, or connect information, which are the core aspects of reasoning [13]. They showed that high-capacity language models can perform a wide range of natural language processing tasks including answering questions, reading comprehension, text summarization, translation, and commonsense reasoning. All these can be achieved in a zero-shot setting without explicit supervision or task-specific training while getting competitive results across many benchmarks [13]. Studies also investigated how chain-of-thought prompting makes reasoning better resulting in performance gains [27]. Prior work also shows how people's information practices are changing in academic settings regarding text summarization and citation generation in this age of LLM [5]. Studies in this area discuss how LLMs can accelerate the literature review process by speeding up traditional approaches including manual search and data synthesis [3, 15]. Studies also discuss the usefulness of LLM-generated data for prototyping and augmentation, and the risks associated with it (e.g., lack of consistency and representativeness) [19]. Besides, they mention how LLMs help in generating interview questions, structuring publications, and reviewing research papers, while also critically analyzing how humans and LLMs are different in their approach and the risks with relying entirely on LLMs [17, 31]. However, to the best of our knowledge, no study has explored how LLMs can guide the generation of threats to validity by producing meaningful, context-specific threats. In this study, we aim to fill this gap based on empirical evidence.

3 Methodology

To answer RQ1 and RQ2, we conducted an empirical experiment to evaluate the quality and novelty of LLM-generated threats by comparing them with the threats reported in the original papers. To answer RQ3, we conducted an iterative taxonomy development process



■ **Figure 1** Overview of the study design for the empirical experiment (RQ1-RQ2) and taxonomy development (RQ3).

to identify and categorize common patterns across the LLM-generated threats. The following subsections describe the experimental setup and threat generation, evaluation framework, and taxonomy development process in detail. Figure 1 presents an overview of the study design.

3.1 Experimental Setup and Threat Generation

To establish a reliable ground truth for evaluating the quality and novelty of the LLM generated threats, we collected 375 research papers from ICSE 2025 [10]. We selected ICSE papers because they represent the influential empirical work by the SE community and operate under consistent constraints. Although journal articles with larger page budgets may provide more comprehensive descriptions of threats to validity, the limited page budget of a conference like ICSE requires authors to prioritize the threats they consider most important, providing a clearer view of actual reporting practices under constraints.

Each paper P was systematically split into two documents automatically: D_M containing the full paper *excluding* the TTV section, and D_{TTV} containing only the TTV section (so $P = D_M \cup D_{TTV}$). This splitting was done to prevent the LLM from directly using the TTV mentioned in the input document and ensure that it generates threats based on the research study’s methodology, results and other information available that helps the model understand the study’s procedure and purpose.

The separated TTV section also works as a reference point and enables us to compare the LLM generated threats against the original threats mentioned in the study and answer whether LLM generated threats overlap with the original threats or have been not reported in the original paper. We selected Gemini 2.5 Flash as the generation model based on its strong reasoning capabilities, its ability to process long documents within a single context window, and its throughput at the scale required to process 375 papers [8, 6]. The temperature was set to zero across all runs to ensure deterministic and reproducible outputs [16].

To investigate whether LLMs can generate study-specific, meaningful threats (RQ1) and identify previously undocumented risks (RQ2), each paper was passed to the model with a structured prompt instructing it to identify the research goal of the study, determine which

validity categories were most relevant to that goal, generate threats that could plausibly affect the study’s conclusions, compare tradeoffs, and recommend appropriate mitigation strategies for each threat. This process produced 2,673 threats across the dataset.

Our approach is retrospective by design, as it generates threats from completed papers. The goal of this preliminary study is not to address validity concerns in already completed studies, but to use completed studies, where both the study methodology and reported TTV are available, to evaluate whether LLMs can identify additional meaningful threats. These findings provide a basis for future work investigating LLM-based threat generation earlier in the research process, when identified threats could inform study design and mitigation decisions.

3.2 Evaluation Framework

To assess the quality of generated threats (RQ1) and their novelty (RQ2), we developed an evaluation framework combining rubric-based scoring with comparison against the threats documented in the original TTV sections. Each of the 2,673 threats generated by the model was scored using a rubric across four dimensions: *relevance*, *specificity*, *clarity*, and *mitigation* where (a) *Relevance* assesses whether the threat meaningfully affects the study’s conclusions, scored as either having impact or no impact, (b) *Specificity* measures how grounded the threat is in the study’s design and context, (c) *Clarity* evaluates how well-articulated and understandable the threat is, and (d) *Mitigation* rates whether the proposed response is appropriate and actionable. The rubric was informed by prior work on empirical validity assessment [20, 7] and LLM output evaluation [26, 21]. To refine the rubric, Gemini was first prompted to generate rating criteria for relevance, specificity, clarity, and mitigation. Then two authors independently graded a subset of the generated threats, compared scores, and adjusted the rubric wherever disagreements revealed unclear or overlapping definitions. This process continued across different batches until consistent scores were achieved across reviewers.

Each generated threat was then automatically compared against the original TTV section of the corresponding paper and classified into one of three categories: (a) present in both the original paper and the model output, (b) present only in the original paper, and (c) generated only by the model. This comparison allowed us to determine how many of the generated threats were previously unreported, and how many were already reported by the authors in the original study.

See the Data Availability section to access all scripts, datasets, and prompts used in this study.

3.3 Taxonomy Development

To answer RQ3, we conducted an iterative taxonomy development process to identify patterns across the generated threats. We grouped the 2,673 threats produced by Gemini into thematic buckets based on semantic similarity. Threats were processed in batches of 100, with the model prompted to generate categories that best captured the similarities among threats within each batch. The resulting categories were compiled and manually reviewed, with overlapping and redundant categories merged for consistency. For example, categories such as “Researcher and Analysis Bias” and “Researcher and Observer Bias” were combined into a single category. This process produced seven final categories: (a) Measurement and Construct Definition Issues, (b) Sampling and Generalizability Issues, (c) Internal Validity and Confounding Factors, (d) Scope and Context Limitations, (e) LLM/AI-Specific

Challenges, (f) Statistical and Conclusion Validity, and (g) Researcher and Analysis Bias. This categorization allowed us to analyze the distribution of each threat category and identify which types of threats were frequently generated by the LLM but not documented in the original TTV sections, leading to a broader coverage of validity concerns.

4 Results

Our analysis of the 2,673 generated threats shows that LLMs are effective at generating relevant, specific, and clearly articulated threats with strong mitigation strategies in most cases, and contribute to novelty by identifying risks not present in the original papers. However, the model did not fully replicate all author-reported threats and occasionally produced generic outputs, positioning LLMs as assistive tools that augment rather than replace human reasoning. The following subsections elaborate on these findings.

4.1 Quality of Generated Threats (RQ1)

We evaluated the quality of generated threats using our rubric across four dimensions: relevance, specificity, clarity, and mitigation. We found that, the model produced high-quality threats across all dimensions. Of the 2,673 generated threats, 2,596 (97%) were classified as having meaningful impact and 77 (3%) had no impact. In terms of specificity, 2,597 threats (97%) were rated as highly specific, 75 (3%) had low specificity, while only 1 was categorized as having medium specificity. Similarly, 2,599 threats (97%) demonstrated high clarity and 74 (3%) showed low clarity. Mitigation quality was slightly lower but still strong: 2,464 threats (92%) received high ratings for concrete, study-specific, and actionable mitigation steps, 130 (5%) mentioned mitigation but lacked depth, and 79 (3%) had generic or boilerplate responses.

To illustrate what these scores mean in practice, we examine a threat generated for the paper “Designing Abandabot: When Does Open Source Dependency Abandonment Matter?”[12] in detail. The paper studies how developers perceive and manage the impact of abandoned dependencies in open source projects. The authors conducted semi-structured need-finding interviews with 22 open-source maintainers. During the interviews, participants discussed how the dependency’s abandonment would impact their project, how the context of their usage of the dependency affects the impact of abandonment, and whether they would want to be made aware of its abandonment and why. The authors also used probing questions when participants had difficulty articulating their reasoning.

The LLM identified social desirability bias as a threat, where participants may provide responses they believe researchers expect rather than their genuine opinions because of feeling implicit pressure to present themselves as proactive or technically informed. This is relevant to the interview-based study design because the findings depend directly on participants’ reported perceptions and reasoning. However, the generated threat did not connect this bias to specific aspects of the interview protocol, such as participants discussing their own dependencies or interacting with the prototype. The model nevertheless correctly identified this as a risk to the study’s internal validity and proposed mitigation strategies including ensuring anonymity and confidentiality, using open-ended questions, and asking probing questions.

Scoring this threat against the rubric reveals an important pattern. While this threat is relevant and clearly articulated, it remains generic and is not strongly grounded in the specific design or findings of the study. Besides, the mitigation strategies, although valid, reflect standard interview practices rather than study-specific interventions. As a result, this

threat is characterized by high clarity but low specificity and low mitigation quality. This example illustrates how our rubric distinguishes lower-quality threats from more contextually grounded threats. Threats that identify valid risks but remain disconnected from the study's specific design, data, and findings tend to score high on relevance and clarity but low on specificity and mitigation, making them lower-quality threats. In contrast, threats that are contextually grounded, i.e., explicitly connected to the study's design, data, and findings, while also being clear and relevant, are considered higher-quality threats that are more specific, practically useful, and provide actionable mitigation.

4.2 Novelty and Coverage of Threats (RQ2)

To analyze the novelty of the generated threats, we compared each generated threat against the original TTV section of the corresponding paper. We found that 61% of the generated threats were new, i.e., they did not overlap with author-reported threats. This implies that the model is able to identify previously undocumented threats, providing an alternative perspective on potential validity concerns that researchers may overlook during the writing process. At the same time, the model did not fully reproduce all author-reported threats, which reflects an important limitation: certain study-specific concerns that authors identify through deep familiarity with their own work are not always recoverable from the paper text alone. This behavior demonstrates the role of LLMs as complementary tools for threat to validity identification. While the model is effective at expanding the space of possible threats, missing some study-specific concerns captured by the original authors indicates that it is not a replacement for human reasoning. The degree of novelty also varied across threat categories, with some categories showing higher rates of new generation than others, which we examine further in RQ3.

4.3 Taxonomy of Threats (RQ3)

We categorized all 2,673 generated threats into seven thematic groups to analyze their distribution across validity dimensions. The largest category was Measurement and Construct Definition Issues (802 threats), followed by Sampling and Generalizability Issues (642 threats) and Internal Validity and Confounding Factors (588 threats). Smaller categories included Scope and Context Limitations (187 threats), LLM/AI-Specific Challenges (160 threats), Statistical and Conclusion Validity (160 threats), and Researcher and Analysis Bias (134 threats). Newly generated threats were prevalent across all categories. Measurement and Construct Definition Issues contained 508 (out of 802) new threats, Internal Validity and Confounding Factors contained 401 (out of 588), and even smaller categories showed high novelty rates, with 107 (out of 134) Researcher and Analysis Bias threats being newly identified. These results show that LLMs generate a diverse set of threats spanning multiple validity dimensions rather than concentrating on a single type. The dominance of certain categories, particularly measurement and sampling concerns, may reflect biases in both model outputs and existing research practices, where these threat types are most commonly discussed and therefore most readily generated. Table 1 summarizes the full distribution of threats across categories and their corresponding novelty proportions.

5 Discussion

This section discusses the implications of our findings for the use of LLMs in validity reasoning and the broader research process. We discuss both the opportunities and limitations of LLM-generated threats, followed by the potential role of LLM in scientific writing, research artifacts, and human-AI research collaboration.

Category	Total	New Threats	% of New
Measurement and Construct Definition	802	508	63%
Sampling and Generalizability	642	321	50%
Internal Validity and Confounding Factors	588	401	68%
Scope and Context Limitations	187	107	57%
LLM/AI-Specific Challenges	160	107	67%
Statistical and Conclusion Validity	160	107	67%
Researcher and Analysis Bias	134	107	80%

■ **Table 1** Distribution of Threats Across Thematic Categories (RQ3).

5.1 Expanding the Space of Validity Concerns

Our findings show the high relevance, clarity, and novelty of LLM generated threats which indicate that LLMs can expand the space of potential validity concerns beyond what authors typically report. This positions LLMs as supportive tools that help researchers identify overlooked risks and develop more comprehensive TTV sections. However, our findings also show that LLM-generated threats are not always perfect, as some were generic, insufficiently grounded in the specific study, or provided generic mitigation strategies, as illustrated by the “Designing Abandabot: When Does Open Source Dependency Abandonment Matter?” [12] example in Section 4.1. Therefore, LLMs do not replace the need for researchers to construct threats to validity, although they can significantly enhance this process. Automated generation also raises questions of responsibility, authorship, hallucinations, and bias, making human oversight essential to validate, refine, and contextualize the outputs. Researchers should verify generated threats instead of accepting them automatically. Rather than fully automating TTV construction, a more appropriate approach is human-in-the-loop collaboration, where researchers iteratively review and adapt LLM-generated suggestions. This can maintain correctness and accountability while benefiting from the model’s breadth.

5.2 From Writing to Co-Writing: Rethinking Scientific Authoring

The ability of LLMs to generate structured and meaningful threats to validity reflects a broader opportunity to support scientific writing. By assisting in identifying gaps, refining arguments, and suggesting mitigations, LLMs can contribute to improving the overall quality and completeness of research papers. This aligns with emerging perspectives in the software engineering community that view AI as a co-authoring assistant capable of enhancing clarity and rigor. Beyond TTV, similar capabilities can support other stages of writing, such as designing methodologies, structuring literature reviews, and framing results. For example, an LLM can suggest missing experimental steps and alternative study designs, or organize prior work into coherent themes, helping researchers move from scattered ideas to a well-structured narrative. When framing results, it can assist in interpreting findings, identifying limitations, and connecting outcomes back to research questions.

An important distinction, however, lies in the type of input required. For generating threats to validity, the model benefits from access to the full context of the paper, including research goals, methodology, and findings, which allows it to reason about potential risks. In contrast, supporting other sections may require different forms of input. Methodology design may rely more on problem framing and prior work, literature review structuring depends on access to a corpus of related papers, and result interpretation requires detailed

experimental outputs. This suggests that effective use of LLMs in scientific writing is not one-size-fits-all, but depends on providing the right contextual grounding for each task. At the same time, reliance on AI-generated content must be carefully managed. Overuse of generic suggestions may reduce the specificity and originality of scientific writing. Over-reliance on automatic threat generation may also reduce researchers' active involvement in thinking about validity concerns and eventually lead to de-skilling. Therefore, the role of LLMs should be framed as augmentative rather than substitutive, supporting researchers in producing better-grounded and more thoughtful contributions. Future work should explore human-in-the-loop approaches that keep researchers actively involved in reviewing, validating, and refining LLM-generated suggestions.

5.3 From Consuming Static Papers to Interacting with Artifacts

LLMs can support not only the writing of research papers but also how they are consumed. While we focus on threats to validity, similar techniques can be extended to other components of a paper to make them more accessible and informative. For instance, LLMs could enable adaptive related work sections that evolve as new research emerges, helping readers stay up to date without manually tracking the literature. They could also suggest alternative analysis perspectives, generate clearer or more engaging visualizations, and provide personalized summaries tailored to different audiences, such as beginners, practitioners, or experts. This points toward a shift from static documents to more dynamic and **interactive research artifacts**. Instead of passively reading a paper, users could engage with it through conversational interfaces, ask follow-up questions, or explore specific sections in more depth. Concepts such as chatting with a paper, automatically generated summaries, or multimodal representations like audio explanations reflect this emerging direction.

However, these possibilities also introduce challenges. AI-generated interpretations may introduce inaccuracies or biases, and over-reliance on such summaries could reduce critical engagement with the original work. Additionally, there is growing pushback from the community regarding the use of AI in research dissemination, including concerns around AI-generated summaries and automated content (for example, see the release of AI summaries in the ACM Digital Library [11]). As a result, balancing accessibility with trust and fidelity remains an important consideration.

5.4 From Retrospective Evaluation to Research Collaboration

LLMs offer meaningful support to researchers well before the writing stage begins, acting as thinking partners across all the stages of the research process. In the earliest phases, they can help sharpen problem framing, surface promising research directions, and identify gaps in existing knowledge. They are especially valuable during literature review, where they can organize large bodies of prior work into coherent themes, reveal connections across studies, and surface relevant papers that a researcher might overlook, supporting more systematic and well-grounded research before it is translated into writing. They can flag missing steps in methodology, propose alternative approaches, and draw attention to assumptions or risks that might otherwise go unnoticed.

However, realizing this potential depends on how AI tools are designed. AI tools for research should be designed not as generic text generators, but as systems tailored to the needs of researchers. A key requirement is support for context-sensitive reasoning, where the quality of outputs depends heavily on the inputs provided. As seen in our study, generating meaningful threats to validity requires access to rich contextual information such

as research goals, methodology, and findings. This demonstrates the importance of designing AI systems that allow researchers to control and curate the context provided to the model. In practice, this implies giving researchers greater autonomy over prompt construction and context engineering. Rather than relying on fixed, opaque prompts, AI tools should expose mechanisms that allow users to iteratively refine inputs, select relevant sections of a paper, and guide the model toward more grounded outputs. Interface support becomes critical here, enabling users to structure inputs, compare outputs across variations, and trace how different contexts influence results. Such design shifts position AI not as a black-box assistant, but as an interactive collaborator which can improve transparency, reproducibility, and trust in AI-assisted research workflows, while also ensuring that researchers remain in control of the reasoning process.

6 Threats to Validity

We organized the threats to validity of our study around construct validity, internal validity, and external validity. We did not consider these threats only after completing the research. During the study design, we considered limitations associated with selecting only ICSE 2025 papers, using one LLM, evaluating generated threats using four dimensions and two raters, and constructing a common taxonomy across different study types. Given the exploratory nature of this study, we intentionally kept the initial scope limited and report these remaining threats transparently. Our findings provide a basis for future work using multiple LLMs, additional raters and repeated runs, journal papers, multiple venues and years, and method-specific validity frameworks. We discuss these threats in detail in the following subsections.

Construct Validity. First, we defined an “undocumented” threat as one generated by the LLM but not identified in the original TTV section of a paper. This does not necessarily mean that the authors overlooked the threat. Authors may have considered the threat and omitted it due to space constraints, mitigated the threat in the method section, or judged it as irrelevant for reporting. Therefore, our comparison measured whether a threat was explicitly documented in the TTV section rather than whether the original researchers were aware of it. Second, our evaluation measured the quality of LLM-generated threats using four dimensions: relevance, specificity, clarity, and mitigation. Although these dimensions provide a structured, checklist-like way to evaluate generated threats, they do not provide a deep assessment of the threats. The rubric was developed and refined by two authors and therefore may carry some subjectivity. Although we iterated until consistent scores were achieved, a larger and more diverse set of raters would strengthen the evaluation. However, the dimensions relevance and specificity in our evaluation rubric helped us to reduce the impact of hallucinated or biased outputs generated by the LLM. They helped assess whether each generated threat is actually relevant and grounded in the study’s design and context. Finally, the taxonomy was also constructed iteratively using the same model that generated the threats, which introduces potential circularity. The resulting categories may reflect patterns in Gemini’s output rather than an independent characterization of the validity threat space. Future work should validate the taxonomy against independently constructed frameworks and through expert review.

Internal and External Validity. Regarding threat to internal validity, the generated threats may depend on the prompt and model configuration used in the study. Although we used the same prompt and set the temperature to zero to reduce output variability, the prompts

used in this study may not be optimal. Alternative prompt formulations or more structured prompting strategies may produce different or higher-quality threats. Repeated runs and systematic evaluation of alternative prompts would provide stronger evidence that our findings are consistent across these design choices.

Regarding threat to external validity, our dataset is limited to ICSE 2025 papers, which may not represent TTV reporting practices across other publication settings. As justified in the Methodology, we selected a single conference and year to maintain consistent publication constraints and keep the scope manageable for this preliminary study. Future work should extend the study to journal papers and papers from multiple venues and years to assess the generalizability of our findings. In addition, papers published in 2025 may have been influenced by AI-assisted writing or proofreading, while older papers could provide a more independent baseline. Future work should therefore examine papers published before the widespread use of generative AI too.

Our evaluation also relies exclusively on Gemini 2.5 Flash. Whether the findings hold for other models (e.g., GPT-4, Claude, Llama) remains an open question. Different models are trained on different data, use different architectures, and undergo different optimization procedures, which can result in variation in their outputs. Therefore, the observed effectiveness, distribution of generated threat types, and degree of study-specific grounding may be specific to the model evaluated in this study. Future work should replicate our evaluation across multiple LLMs and repeated runs to assess the generalizability and consistency of the findings.

Finally, our study combines papers that use different empirical research methods, such as experiments, case studies, surveys, and mining software repositories, which may involve different validity concerns. We used a common taxonomy because, as an initial exploration, our goal was to examine the overall landscape of LLM-generated threats and understand how LLMs can support threat generation across empirical SE research. We do not claim that this taxonomy is universally applicable across all research methods. Future work should investigate method-specific threats and taxonomies based on existing validity guidelines for individual research methods.

7 Conclusion

This study investigated whether large language models can generate study-specific, meaningful, and novel threats to validity for empirical software engineering research. Across 375 papers from ICSE 2025, the pipeline produced 2,673 threats, of which 97% were rated as relevant, specific, and clearly articulated, 92% received high ratings for mitigation quality, and 61% were absent from the original papers, indicating the model consistently identified concerns the authors had not reported. The high novelty rates in categories such as Researcher and Analysis Bias (80%), Internal Validity and Confounding Factors (68%), and LLM/AI-Specific Challenges (67%) suggest these are the threat types most commonly missing from author-reported sections. The qualitative analysis further shows that threat quality depends directly on how well the model grounds its output in the study's specific design, data, and findings, with contextually grounded threats scoring higher across all four rubric dimensions.

The results suggest that LLMs are best understood as thinking partners for validity reasoning, capable of expanding the space of concerns a researcher considers without substituting the domain expertise and contextual familiarity that only the author can bring. Though the model does not fully reproduce all concerns that the authors identify through deep familiarity with their own work, the evidence suggests that it can meaningfully raise the quality and completeness of validity sections across the field.

Future work should extend our approach to other venues and model families, explore human-in-the-loop refinement workflows to improve specificity and actionability, and investigate whether LLMs can be applied proactively during study design rather than retrospectively applying them after writing, while shifting validity reasoning from a secondary concern and afterthought to a deliberately embedded activity throughout the research process.

Data Availability Statement

All data and scripts supporting this study, including the 375 ICSE input papers, the threat generation prompt, the generated threats dataset, all evaluation scripts, and the scoring sample, are publicly available in our replication package on Zenodo at <https://doi.org/10.5281/zenodo.20453824> under a CC-BY 4.0 license.

References

- 1 Rand Alchokr, Evelyn Starzew, Gunter Saake, Thomas Leich, and Jacob Krüger. The impact of ai language models on scientific writing and scientific peer reviews: A systematic literature review. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, pages 1–6, 2024. doi:10.1145/3677389.3702508.
- 2 Apostolos Ampatzoglou, Stamatia Bibi, Paris Avgeriou, and Alexander Chatzigeorgiou. Guidelines for managing threats to validity of secondary studies in software engineering. In *Contemporary empirical methods in software engineering*, pages 415–441. Springer, 2020. doi:10.1007/978-3-030-32489-6_15.
- 3 Shouvik Ahmed Antu, Haiyan Chen, and Cindy K Richards. Using llm (large language model) to improve efficiency in literature review for undergraduate research. *Llm@ Aied*, pages 8–16, 2023. URL: <https://ceur-ws.org/Vol-3487/short2.pdf>.
- 4 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 5 Gobinda Chowdhury and Sudatta Chowdhury. Ai-and llm-driven search tools: A paradigm shift in information access for education and research. *Journal of Information Science*, page 01655515241284046, 2024.
- 6 Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint*, 2025. arXiv:2507.06261.
- 7 Robert Feldt and Ana Magazinius. Validity threats in empirical software engineering research—an initial survey. In *Proceedings of the 22nd International Conference on Software Engineering & Knowledge Engineering (SEKE)*, pages 374–379, 2010.
- 8 Google DeepMind. Gemini 2.5 flash, 2025. URL: <https://deepmind.google/technologies/gemini/flash>.
- 9 Patricia Lago, Per Runeson, Qunying Song, and Roberto Verdecchia. Threats to validity in software engineering—hypocritical paper section or essential analysis? In *Proceedings of the 18th ACM/IEEE International symposium on empirical software engineering and measurement*, pages 314–324, 2024. doi:10.1145/3674805.3686691.
- 10 David Lo, Corina Pasareanu, Timothy C. Lethbridge, and Lionel Briand, editors. *ICSE ’25: Proceedings of the IEEE/ACM 47th International Conference on Software Engineering*. IEEE Press, 2025. doi:10.1109/ICSE55347.2025.
- 11 Anil Madhavapeddy. Dear ACM, you’re doing AI wrong but you can still get it right, 2025. URL: <http://doi.org/10.59350/c84g4-5zt58>.

- 12 Courtney Miller, Hao He, Weigen Chen, Elizabeth Lin, Chenyang Yang, Bogdan Vasilescu, and Christian Kästner. Designing abandabot: When does open source dependency abandonment matter? In *Proc. Int'l Conf. Software Engineering (ICSE)*, 2026.
- 13 Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- 14 Paul Ralph and Ewan Tempero. Construct validity in software engineering research and software metrics. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*, pages 13–23, 2018. doi:10.1145/3210459.3210461.
- 15 Nitin Liladhar Rane, Abhijeet Tawde, Saurabh P Choudhary, and Jayesh Rane. Contribution and performance of chatgpt and other large language models (llm) for scientific and research advancements: a double-edged sword. *International Research Journal of Modernization in Engineering Technology and Science*, 5(10):875–899, 2023.
- 16 Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 7346–7356, 2024. doi:10.18653/V1/2024.FINDINGS-EMNLP.432.
- 17 John Roberts, Max Baker, and Jane Andrew. Artificial intelligence and qualitative research: The promise and perils of large language model (llm)‘assistance’. *Critical Perspectives on Accounting*, 99:102722, 2024.
- 18 Aaisha AL Roshdi, Fareed Awae, and Ahmad Yussuf. The impact of artificial intelligence tools on developing and improving academic writing. *Digital Technologies Research and Applications*, pages 179–197, 2026.
- 19 Luca Rossi, Katherine Harrison, and Irina Shklovski. The problems of llm-generated data in social science research. *Sociologica: International Journal for Sociological Debate*, 18(2):145–168, 2024.
- 20 Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131–164, 2009. doi:10.1007/S10664-008-9102-8.
- 21 Sherry Shi, Renyao Wei, Michele Tufano, José Cambronero, Runxiang Cheng, Franjo Ivančić, and Pat Rondon. Towards a human-in-the-loop framework for reliable patch evaluation using an llm-as-a-judge. *arXiv preprint*, 2025. arXiv:2511.10865.
- 22 Janet Siegmund, Norbert Siegmund, and Sven Apel. Views on internal and external validity in empirical software engineering. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 1, pages 9–19. IEEE, 2015. doi:10.1109/ICSE.2015.24.
- 23 Dag IK Sjøberg and Gunnar R Bergersen. Improving the reporting of threats to construct validity. In *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering*, pages 205–209, 2023.
- 24 Marion K Slack and Jolaine R Draugalis Jr. Establishing the internal and external validity of experimental studies. *American journal of health-system pharmacy*, 58(22):2173–2181, 2001.
- 25 Roberto Verdecchia, Emelie Engström, Patricia Lago, Per Runeson, and Qunying Song. Threats to validity in software engineering research: A critical reflection. *Information and Software Technology*, 164:107329, 2023. doi:10.1016/J.INFSOF.2023.107329.
- 26 Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of llm-as-a-judge in llm alignment tasks: Explainable metrics and diverse prompt templates. *arXiv preprint*, 2024. arXiv:2408.13006.
- 27 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- 28 Claes Wohlin, Martin Höst, and Kennet Henningsson. Empirical research methods in software engineering. In *Empirical methods and studies in software engineering: Experiences from ESERNET*, pages 7–23. Springer, 2003. doi:10.1007/978-3-540-45143-3_2.
- 29 Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, Anders Wesslén, et al. *Experimentation in software engineering*, volume 236. Springer, 2012.

- 30 Hyrum K Wright, Miryung Kim, and Dewayne E Perry. Validity concerns in software engineering research. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*, pages 411–414, 2010. doi:10.1145/1882362.1882446.
- 31 Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024. URL: <https://aclanthology.org/2024.lrec-main.816>.