International Journal of Research and Review
[Indexed, Peer-Reviewed Journal]
Volume 12; Issue: 9; September 2025
Website: www.ijrrjournal.com
Review Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Human-AI Collaboration in Validating and Refining LLM Summaries of Test Automation Results

Alex Thomas Thomas

Saransh Inc, New Jersey, USA

Corresponding Author: Alex Thomas Thomas

DOI: https://doi.org/10.52403/ijrr.20250941

ABSTRACT

With Large Language Models (LLMs) persistently demonstrating the capabilities of automating software test procedures, the need for effective human monitoring and cooperation in confirming AI-produced test summaries has grown to become a top priority in software quality assurance. This integrative review synthesizes current evidence for human-AI collaborative frameworks for verifying and enhancing LLM-generated summaries of test automation reports, investigating the intersection of artificial intelligence strengths and human competence in ensuring reliable software testing results. The review systematically reviews existing methodological paradigms for human-in-the-loop testing protocols, with particular emphasis on quality engineering techniques instilling confidence and safety in software systems that utilize LLMs, and empirical research comparing LLM and human judge performance in software engineering contexts to investigate the potential and potential constraints of AI systems as standalone judges of test quality. Underlying the analysis here is a look at collaborative intelligence frameworks that integrate human knowledge and AI potential smoothly within software testing scenarios, evaluating structured interaction designs that enable significant human-LLM collaboration and examining functionality-aware decision-making that can maximize the dependability of AI-generated test summaries through knowledge-validation mechanisms that integrate LLMs with human overseeing mechanisms. Discoveries emphasize that strong human-AI collaboration in testing result validation requires well-designed interaction paradigms leveraging human domain expertise while exploiting AI processing abilities, with trust and transparency being key determinants of establishing including robust evaluation metrics, mixed-methods validation schemes, and human-centric approaches to AI-generated technical documentation. The review finishes by proposing a harmonized methodological approach for measuring human-AI collaboration effectiveness in test automation cases, highlighting the necessity of systematic validation processes, interpretable AI decision-making algorithms, and continuous human oversight in guaranteeing software quality requirements, with significant implications for software engineering practitioners seeking to integrate LLM functionality into existing testing procedures while maintaining high-quality assurance standards.

Keywords: Human-AI collaboration, Large Language Models, test automation, software testing, validation frameworks, quality assurance, human-in-the-loop systems, collaborative intelligence

INTRODUCTION

The emergence of Large Language Models (LLMs) in recent times has brought about prospects for automating software testing workflows, such as the synthesis of comprehensive summaries of complex test automation results [3]. However, while LLMs demonstrate astonishing capabilities in processing and integrating large volumes of testing data, their output must be carefully verified for correctness, completeness, and utility for software development teams [2]. This difficulty has provided the impetus for growing fascination with human-AI collaborative systems that leverage both the efficiency of automated systems and the critical judgment of human specialists to generate more reliable and trustworthy test result summaries [1],[4].

Recent research has emphasized the necessity of formulating systematic approaches to validating AI-generated content for software engineering use cases, particularly in relation to human oversight for critical decision-making [15],[13]. Researchers have discovered that effective human-AI collaboration in the validation of test outcomes can not only improve the quality of generated summaries but also facilitate trust among development teams who utilize the insights in making crucial software quality decisions [12],[13]. The intersection of human judgment and LLM capabilities presents an opportunity to create more robust testing pipelines that wed the scalability of auto-mated evaluation with the contextual understanding and domain knowledge that human reviewers provide to the validation process [14],[15],[1].

OVERVIEW OF LLMs IN SOFTWARE TESTING

Large Language Models have revolutionized software testing by introducing mature capabilities from the initial generation of test cases to the end analysis of results and reporting [3]. Large Language Models demonstrate unparalleled ability to understand natural language requirements and translate them into executable test scenarios perfectly bridging the gap between business specs and technical realization [11]. LLMs excel where there is a requirement for intricate reasoning and pattern recognition, such as in the case of mobile GUI testing where they can make functionality-aware decisions that are highly reflective of human thought patterns [11]. The ability of the models to handle vast quantities of test data and supply clear, contextually relevant summaries has been highly beneficial for businesses looking to ramp up their testing without sacrificing on coverage [8],[6]. Moreover, LLMs have also been shown to perform exceptionally well in test-driven interactive code generation, where they dynamically produce and modify test cases with real-time feedback as well as shifting requirements [8].

The technical capability of LLMs extends beyond minimal automation to include sophisticated analysis and interpretation of test outcomes. The models can effectively synthesize complex test outcomes, identify failure patterns, and generate comprehensive technical reports to support de-bugging and quality assurance tasks [13]. Research has established that LLMs are potential sage arbiters in software development environments, passing judgment on code quality, test coverage, and even judging the quality of other AI-created content [2]. Empirical research, however, presents enormous limitations in their ability to completely replace human evaluators, particularly in scenarios that involve deep domain understanding, contextual awareness, and acute judgments [2],[15]. The models are prone to do poorly on edge cases, may be biased in their rulings, and can generate sounding reasonable but factually incorrect summaries, underscoring the imperative requirement for human reviewing and vetting [15],[1].

The evolution of LLMs has come in software testing to design hybrid models that draw on the strengths of artificial and

human abilities. These joint models tend to position LLMs as powerful auxiliaries responsible for doing mundane analysis, pattern finding, and initial summary generation, with human experts holding central positions in validation, context understanding, and strategic choices [12],[13]. Research has indicated that such human-AI cooperation not only increases the dependability and accuracy of test results but also improves confidence and assurance in AI-based testing processes [15],[1]. Most effective deployments entail rigorously designed interactive schemes with well-defined roles and responsibilities of the human and AI elements, respectively, and ensure key testing judgments are maintained under human control and supported by AI strengths for efficiency and scalability [4],[6]

Table1: Human vs. LLM Evaluators

Evaluation Criteria	Human Evaluators	LLMs
Contextual Understanding	High	Limited
Bias Handling	Varies	Prone to systematic bias
Efficiency	Lower	High
Handling Edge Cases	Reliable	Struggles
Consistency	Variable	High, but not always accurate

HUMAN-AI COLLABRATION IN TEST SUMMARY VALIDATION

Human-AI collaboration in test summary validation is a key paradigm that combines the computational power of Large Language Models with human knowledge and critical thinking of evaluators [14],[15]. This collaborative paradigm recognizes that while LLMs can process vast amounts of test data at speed and generate lengthy summaries, human oversight remains essential to ensure accuracy, relevance, and usability of such outputs [13],[1]. It has been proven that effective validation frameworks typically comprise well-disciplined interaction patterns whereby LLMs perform preliminary data processing and abridgment generation, while hu-man experts conduct important review, context understanding, and quality assurance [12],[13]. The collaborative model allows hu-man validators to focus their skills on high-level decision-making and edge case analysis, rather than frittering away time on routine data processing tasks that can be effectively performed by AI systems [1].

Human-AI collaborative work validation must be treated with care in terms of trust, transparency, and methodological rigor to yield reliable results [15],[1]. Studies have indicated that good collaboration models entail multi-level validation through consistency checks through automation, expert human validation, and improvement loop iterations to allow continuous refinement of summary quality [15],[6]. The human's contribution in this partnership extends beyond the simple acceptance or rejection of AI-generated content to being involved in refining, enhancing, and situating the summaries within project-specific background and domain knowledge [14],[1]. It is empirically established that if appropriately framed, such collaborative approaches not only improve the quality and validity of test summaries but also improve the confidence of development teams using these results for making important decisions about software quality [1],[4].

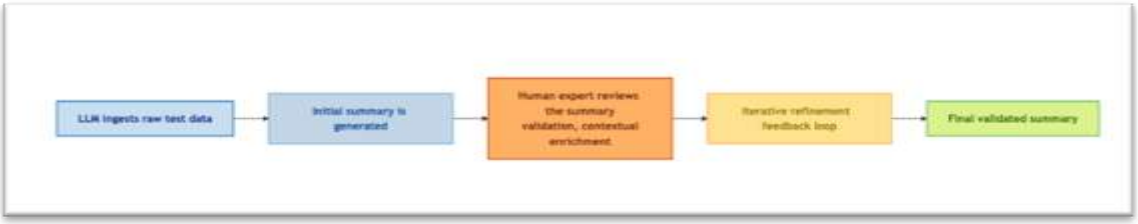

Fig. 1: Human-AI Collaboration Workflow in Test Summary Validation

METHODS FOR GENERATING LLM-BASED TEST SUMMARIES

Large Language Models employ various sophisticated techniques to generate comprehensive test summaries out of automation results, applying their natural language processing abilities to transform raw test data into valuable information [3],[13]. Their primary approach is ordered data ingestion wherein LLMs ingest various inputs of testing data like execution logs, failure reports, performance metrics, and coverage metrics and create readable narrative summaries [14],[1]. Such models utilize advanced pattern discovery techniques to identify key test outcomes, categorize failures in severity and class, and reap key performance measures of highest concern to development teams [6]. The generation process usually entails multi-step analysis whereby LLMs first break down and interpret the technical data, later combining this data into readable forms that highlight key trends, outliers, and actionable suggestions [13],[3].

The test summary generation methodology using LLM has some re-fining mechanisms in place to ensure accuracy and pertinence of output [8],[11]. Real-time feedback and iterative refinement are facilitated by interactive generation methods, where the LLM has the capability to re-tune its summary focus based on user needs and project requirements [8],[14]. Advanced techniques incorporate functionality-aware decision-making, where LLMs account for the contextually important nature of different test components and prioritize information accordingly [11]. Such frameworks often employ template-based models with dynamic content generation, ensuring consistency of form while ensuring content flexibility based on the specific quality of each test case [3],[1]. Incorporating domain knowledge and testing best practices in the generation process ensures LLMs in producing summaries that are not only representative of the technical results but also strategic in terms of directing software quality improvement [16],[13].

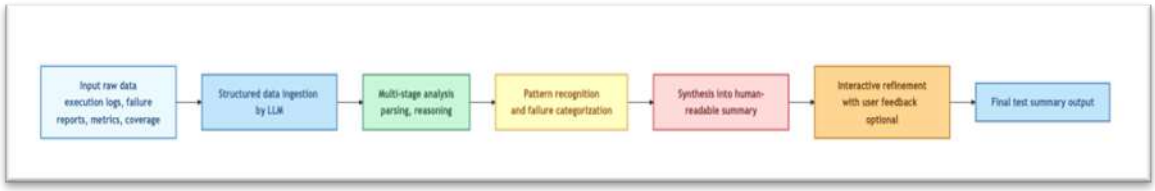

Fig 2: LLM-Based Test Summary Generation Process

ROLE OF HUMAN FEEDBACK IN REFINING SUMMARIES

Human feedback is essential to enhance the precision and validity of LLM-produced test summaries by progressively enhancing them with the help of domain expertise and contextual understanding [14],[1]. The feedback mechanism allows human experts to identify errors, coverage lacunae, and misinterpretations in AI-generated summaries and provide specific recommendations for improvement beyond simple rejection or approval [6],[13]. Literature indicates that human feedback mechanisms enable the continuous learning and adaptation of LLM outputs, with human experts being capable of pinpointing critical test results that may have been overlooked, correcting technical misunderstandings, and mapping summaries to project-based quality standards and business objectives [1],[10]. This human-in-the-loop approach transforms the summary generation process from one automated process into a collaborative refinement cycle that iteratively improves both content relevance and quality [14],[1].

The use of human feedback to optimize LLM summaries involves multiple dimensions of evaluation and optimization, including technical validity, contextual relevance, and strategic value in decision-

making [6],[15]. Human evaluators provide essential feedback on the instrumental utility of machine summaries, highlighting areas where greater specificity is required or where technical terminology needs to be resolved into terms accessible to the general knowledge of stakeholders [13],[6]. Experiments have shown that effective feedback models incorporate explicit correction as well as implicit guidance from user interactions to allow LLMs to learn from patterns of expert judgment and modify their future output accordingly [8],[10]. The feedback loop also addresses concerns of transparency and trust by enabling human experts to validate important discoveries, validate the logical coherence of summary narratives, and guarantee that important edge cases or unusual results receive appropriate attention in the ultimate output [15],[6].

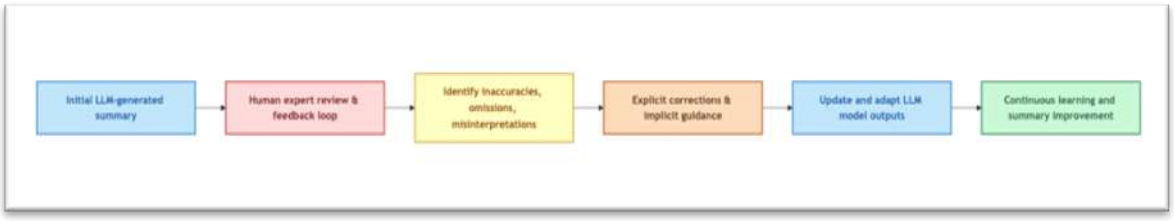

Fig 3: Human Feedback Integration in LLM Summary Refinement

EVALUATION TECHNIQUES AND METRICS

The evaluation of LLM-generated test summaries in human-AI collaborative settings involves a complex set of approaches that analyze both technical accuracy and practical usefulness based on multiple measurement techniques [16],[15]. Quantitative evaluation approaches include accuracy metrics that measure the validity of technical information, completeness metrics that ascertain coverage of major test results, and consistency metrics that evaluate the coherence and logical flow of generated summaries [2],[6]. Human judgment-based methods involve evaluating expert assessments with structured rubrics against dimensions such as readability, actionability, and faithfulness to stakeholder needs [6],[1]. Mixed-methods appraisal frameworks combine automated measures with qualitative human assessments, providing a unifying view of summary quality that encompasses both objective technical correctness and subjective usability issues [15],[6]. These evaluation techniques generally include comparative comparison with human- and AI-generated summaries to yield baseline performance metrics and identify improvement areas [2],[14].

Advanced evaluation methods emphasize measuring human-AI collaboration effectiveness itself. They compare how well the hybrid system works compared to purely human or purely AI approaches [6],[15]. Trust and transparency metrics assess the faith users place in AI-generated summaries and the perceived trustworthiness of the collaborative validation process [15],[1]. Interactive evaluation measures the efficacy and productivity of feedback loops, e.g., response time to human feedback, adaptation quality, and learning retention across multiple refinement iterations [8],[10]. Custom metrics to evaluate the collaborative process have also been investigated, e.g., cognitive load analysis for human annotators, time-to-validation metrics, and quality improvement across subsequent iterations [14],[1]. These comprehensive evaluation frameworks ensure that both collaborative process efficiency and technical quality of summaries are continuously under monitoring and optimization to maintain high degrees of accuracy and usability in production environments [6],[4].

Table 2: Comparative Evaluation Matrix of Testing Approaches

Approach	Accuracy	Scalability	Interpretability	Bias	Adaptability
LLM-only	Moderate (subject to errors, especially on edge cases)	High (can process large volumes quickly)	Low (black-box reasoning)	High (can reflect training data biases)	Moderate (limited to training data, no contextual learning)
Human-only	High (deep domain expertise)	Low (time-consuming and resource-intensive)	High (transparent reasoning)	Low (judgment mitigates bias)	High (can adapt to new contexts)
Human-AI collaboration	High (synergistic combination)	Moderate-High (AI enhances speed, humans ensure quality)	Moderate-High (human interprets AI outputs)	Reduced (human oversight mitigates AI biases)	High (interactive learning and adaptation)

PRACTICAL USE CASES AND EXAMPLES

Real-world use cases of human-AI collaboration for validation of test summaries demonstrate significant real-world value in numerous software development contexts, particularly large enterprise software where extensive test result examination is necessary [8],[11]. GUI testing of mobile applications is an interesting use case where LLMs generate first-order summaries of automated test executions, which subsequently receive verification from human experts assuring the functional correctness and user experience effects of identified defects [11],[14]. In the context of continuous integration, collaborative workflows have worked smoothly in which LLMs analyze overnight test runs and produce initial summaries that development teams review and improve during morning standup meetings to prioritize bug fixing and feature work [3],[13]. Interactive code generation situations illustrate another practical use where LLMs create testing summaries that are iteratively refined by coders in feedback cycles, resulting in better and more actionable test reports [8],[12].

Industry research includes diverse patterns of adoption that tune preferences of human-AI collaboration to organizational needs and technological constraints [13],[6]. Large software companies have had multi-level verification systems in place where LLMs generate unique summaries of technical content for QA teams, which are followed by human experts generating executive summaries for stakeholder communication [13],[1]. Research laboratories have built collaborative structures for the verification of findings from testing of sophisticated systems, with subject matter experts working alongside AI systems such that major edge cases and performance issues receive proper attention in report-outs [6],[5]. Knowledge graph validation projects show how human-AI collaboration extends traditional software testing to include data quality verification, with specialists verifying AI-generated reports on data integrity tests and schema validation results [5]. These real-world implementations consistently show that collaborative processes maintain superior levels of accuracy and higher levels of stakeholder satisfaction compared to single-automated or single-manual summary generation processes [6],[13].

CHALLENGES IN TRUST AND RELIABILITY

Trust and dependability issues present significant difficulties in human-AI collaboration for test summary validation, primarily from the inherent uncertainty and "black box" nature of LLM decision-making processes [15],[1]. Human evaluators often struggle with establishing appropriate levels of trust in AI-produced summaries due to the "black box" nature of LLMs, in which the justification for specific conclusions or

exclusions is not discernible [15],[7]. This transparency issue is compounded by the fact that LLMs can generate plausible-sounding but in-fact incorrect summaries, and human beings are unable to identify slight inaccuracies without needing extensive domain knowledge and tedious verification mechanisms [2],[16]. Research indicates that trust calibration becomes a unique concern when LLMs perform inconsistently on various types of test cases in such a way that one is skeptical of when human oversight is required most [15],[4].

The threat to reliability extends beyond individual summary accuracy to encompass systematic biases and restriction that can contaminate the whole collaborative process [2],[9]. LLMs may possess accustomed blind spots in specific technical domains or consistently misread classes of test failures, creating patterns of unreliability that will require human evaluators to continue monitoring [2],[16]. Studies suggest that even experienced software testing professionals will end up in either over-reliance on AI output or too much skepticism, both of which reduce the effectiveness of the collaborative validation process [15],[1]. Moreover, the evolutionary nature of software systems means that LLM performance will ultimately decay with time since test conditions change constantly, which requires ongoing recalibration of trust relationships and validation processes to provide reliable outputs [1],[4]. These challenges call for the design of sophisticated frameworks that can systematically analyze and report AI reliability and equip human evaluators with the tools and information needed to make well-informed trust judgments [9],[15].

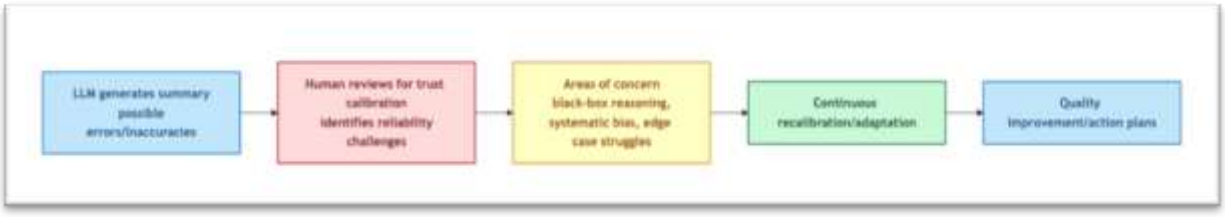

Fig 3: Trust and Reliability Challenges in Human-AI Collaboration

FUTURE DIRECTIONS

Adaptive Collaboration: Learning from Human Feedback

The creation of adaptive collaboration systems is a chief frontier in human-AI collaboration, whereby LLMs learn incrementally from human feedback patterns to improve their summary generation and validation processes over time [8],[10]. Designed human-LLM interaction structures have been demonstrated to reveal exploration and exploitation phenomena through which AI systems adapt their behavior based on aggregated feedback from domain experts [10],[1]. Future work focuses on developing sophisticated feedback learning mechanisms that allow LLMs to detect common patterns of correction, understand context-specific preferences, and automatically tailor their summary generation approach to more closely align with human expert expectations [1],[4]. These adaptive systems will incorporate reinforcement learning techniques that enable continuous improvement of collaboration quality, with human feedback being employed as a training signal for both technical correctness and communication style of generated summaries [8],[6].

Multi-agent and Swarm-human Hybrid Approaches

Current research examines multi-agent collaborative systems that bring together diverse AI systems and human experts with the goal of creating more robust and inclusive validation procedures [4],[9]. These approaches merge the complementary strengths of different AI models with the

maintenance of essential human oversight and decision-making authority over critical validation tasks [9],[5]. Future rollouts will also see swarm intelligence principles applied with multiple LLMs collaborating with human experts in distributed verification networks, enabling parallel processing of complex test results while maintaining quality through human coordination [4],[12]. Additionally, advances in evaluation frameworks and benchmarks are driving the development of standardized metrics and methodologies for assessing human-AI collaboration effectiveness, including comprehensive benchmarks that measure not only technical accuracy but also collaboration efficiency, trust calibration, and long-term learning outcomes [6],[15]. These standardized evaluation approaches will facilitate better comparison of different collaborative approaches and enable systematic optimization of human-AI workflows across diverse software testing contexts [6],[2].

CONCLUSION

The existing evidence invariably shows that human judgment is even still in-dispensable for the achievement of confident test automation, particularly when validating and fine-tuning summaries generated by LLM, as while LLMs have shown incredible competence in handling gigantic amounts of testing data and generating readable summaries, their incapacity to attain contextual under-standing, domain-specific facts, and critical reasoning means that it is required to retain human intervention [2],[15],[6]. Studies consistently illustrate that LLMs can generate plausible but incorrect abstractions, reflect systematic bias, and fare badly on boundary cases where nuanced interpretation is called for, reaffirming that the collaborative effort hybrid mode combining AI efficiency with human intelligence is superior to exclusive automation or exclusive manual endeavor [15],[1],[12],[13]. Follow-up research should focus on developing adaptive learning systems that can better absorb human feedback patterns and optimize collaboration efficiency with time, and some of the key areas for research should be standardized test platforms that can comprehensively quantify technical accuracy and collaboration efficiency, multi-agent solutions that leverage diverse AI capabilities and maintain humans in control, and trust calibration technology that supports humans in making appropriate reliance judgments [8],[10],[6],[4],[9]. Industry adoption will be facilitated through best practice human-AI collaborative workflow development, training initiatives to facilitate successful human and AI system collaboration, and organization structures that facilitate sustainable collaboration models, leading to the establishment of mature collaborative frameworks that take advantage of human and artificial intelligence strengths while addressing their respective limitations in core software quality assurance activities [1],[13],[6].

Declaration by Authors

Acknowledgement: None
Source of Funding: None
Conflict of Interest: No conflicts of interest declared.

REFERENCES

1. Kathiresan, G., 2025. Human-in-the-Loop Testing for LLM-Integrated Software: A Quality Engineering Framework for Trust and Safety. *Authorea Preprints*
2. Wang, R., Guo, J., Gao, C., Fan, G., Chong, C.Y. and Xia, X., 2025. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering*, *2*(ISSTA), pp.1955-1977.
3. Sherifi, B., Slhoub, K. and Nembhard, F., 2024. The potential of llms in automating software testing: From generation to reporting. *arXiv preprint arXiv:2501.00217*.
4. Ronanki, K., 2025, June. Facilitating Trustworthy Human-Agent Collaboration in LLM-based Multi-Agent System Oriented Software Engineering. In *Proceedings of the 33rd ACM International Conference on the*

Foundations of Software Engineering (pp. 1333-1337).
5. Tsaneva, S., Dessì, D., Osborne, F. and Sabou, M., 2025. Knowledge graph validation by integrating LLMs and human-in-the-loop. *Information Processing & Management*, *62*(5), p.104145.
6. Fragiadakis, G., Diou, C., Kousiouris, G. and Nikolaidou, M., 2024. Evaluating human-ai collaboration: A review and methodological framework. *arXiv preprint arXiv:2407.19098*.
7. Raikov, A., Giretti, A., Pirani, M., Spalazzi, L. and Guo, M., 2024. Accelerating human–computer interaction through convergent conditions for LLM explanation. *Frontiers in Artificial Intelligence*, *7*, p.1406773
8. Fakhoury, S., Naik, A., Sakkas, G., Chakraborty, S. and Lahiri, S.K., 2024. Llm-based test-driven interactive code generation: User study and empirical evaluation. *IEEE Transactions on Software Engineering*.
9. Rastogi, C., Tulio Ribeiro, M., King, N., Nori, H. and Amershi, S., 2023, August. Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 913-926)
10. Flores Romero, P., Fung, K.N.N., Rong, G. and Cowley, B.U., 2025. Structured human-LLM interaction design reveals exploration and exploitation dynamics in higher education content generation. *npj Science of Learning*, *10*(1), p.40.
11. Liu, Z., Chen, C., Wang, J., Chen, M., Wu, B., Che, X., Wang, D. and Wang, Q., 2024, April. Make llm a testing expert: Bringing human-like interaction to mobile gui testing via functionality-aware decisions. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (pp. 1-13).
12. Chen, N., HuiKai, A.L., Wu, J., Hou, J., Zhang, Z., Wang, Q., Wang, X. and He, B., 2025. XtraGPT: LLMs for Human-AI Collaboration on Controllable Academic Paper Revision. *arXiv preprint arXiv:2505.11336*.
13. Ricca, F., Marchetto, A. and Stocco, A., 2025. A multi-year grey literature review on AI-assisted test automation. *Information and Software Technology*, p.107799.
14. Gao, Q., Xu, W., Pan, H., Shen, M. and Gao, Z., 2025. Human-Centered Human-AI Collaboration (HCHAC). *arXiv preprint arXiv:2505.22477*.
15. Wang, Q., Wang, J., Li, M., Wang, Y. and Liu, Z., 2024. A roadmap for software testing in open collaborative development environments. *arXiv preprint arXiv:2406.05438*.

How to cite this article: Alex Thomas Thomas. Human-AI collaboration in validating and refining LLM summaries of test automation results. *International Journal of Research and Review*. 2025; 12(9): 393-401. DOI: *https://doi.org/10.52403/ijrr.20250941*

******