UNIVERSITAS
GADJAH MADA
Rancang Bangun Sistem Pembangkit Skenario Pengujian Multi-Persona Berbasis Large Language Model (LLM)
Tsania Qurrota A'yunin Qulub, Divi Galih Prasetyo Putri, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR PUSTAKA

[1] J. Wang, Y. Huang, C. Chen, Z. Liu, S. Wang, and Q. Wang, "Software testing with large language models: Survey, landscape, and vision," *IEEE Transactions on Software Engineering*, 2024.

[2] International Software Testing Qualifications Board, "Certified tester foundation level syllabus," https://istqb.org/wp-content/uploads/2024/11/ISTQB_CTFL_Syllabus_v4.0.1.pdf, 2024, version 4.0.1.

[3] Z. Yang, R. Huang, C. Cui, N. Niu, and D. Towey, "Requirements-based test generation: A comprehensive survey," *ACM Transactions on Software Engineering and Methodology*, 2025.

[4] S. Akbarova, F. Dobslaw, and R. Feldt, "Understanding on the edge: Llm-generated boundary test explanations," in *Proceedings of the 7th ACM/IEEE International Conference on Automation of Software Test (AST 2026)*, 2026.

[5] J. Fischbach, J. Frattini, A. Vogelsang, D. Mendez, M. Unterkalmsteiner, A. Wehrle, P. Restrepo Henao, P. Yousefi, T. Juricic, J. Radduenz, and C. Wiecher, "Automatic creation of acceptance tests by extracting conditionals from requirements: Nlp approach and case study," *Journal of Systems and Software*, vol. 197, p. 111549, 2023.

[6] S. U. Singh and A. S. Namin, "A survey on chatbots and large language models: Testing and evaluation techniques," *Natural Language Processing Journal*, vol. 10, 2025.

[7] X. Li, Y. Hao, J. Hou, J. Huang, Q. Wen, S. Huang, Y. Liu, X. Liu, Y. Fan, Y. Wang *et al.*, "Matraix: Simulating the world with 8.3 billion persona agents," *arXiv*, 2026.

[8] D. Amin, J. Salminen, F. Ahmed, S. M. H. Tervola, S. Sethi, and B. J. Jansen, "Creating and evaluating personas using generative ai: A scoping review of 81 articles," in *Proceedings of CHI 2026*, 2026.

[9] B. Chu *et al.*, "Large language models for unit test generation: Achievements, challenges, and opportunities," *arXiv*, 2025.

UNIVERSITAS GADJAH MADA
Rancang Bangun Sistem Pembangkit Skenario Pengujian Multi-Persona Berbasis Large Language Model (LLM)
Tsania Qurrota A'yunin Qulub, Divi Galih Prasetyo Putri, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

[10] A. Celik and Q. H. Mahmoud, "A review of large language models for automated test case generation," *Machine Learning and Knowledge Extraction*, vol. 7, no. 3, 2025.

[11] M. D. Santana, C. Magalhães, and R. d. S. Santos, "Software testing with large language models: An interview study with practitioners," *arXiv*, 2025.

[12] VoidZero Inc. and Vite Contributors, "Vite: Next generation frontend tooling," https://vite.dev/guide/, 2026.

[13] Microsoft, "The typescript handbook," https://www.typescriptlang.org/, 2026.

[14] Hono, "Hono, a web framework built on web standards," https://hono.dev/, 2026.

[15] K. P. Gaffney, M. Prammer, L. Brasfield, D. R. Hipp, D. Kennedy, and J. M. Patel, "Sqlite: Past, present, and future," *Proceedings of the VLDB Endowment*, vol. 15, no. 12, pp. 3535–3547, 2022.

[16] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, "Constitutional ai: Harmlessness from ai feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[17] A. Biørn-Hansen, T.-M. Grønli, and G. Ghinea, "A survey and taxonomy of core concepts and research challenges in cross-platform mobile development," *ACM Computing Surveys*, vol. 51, 2018.

[18] B. R. Korraprolu, P. Pinninti, and Y. R. Reddy, "Test case generation for requirements in natural language: An llm comparison study," in *Proceedings of the 18th Innovations in Software Engineering Conference (ISEC 2025)*. ACM, 2025.

[19] A. Guzu, G. Nicolae, H. Cucu, and C. Burileanu, "Large language models for c test case generation: A comparative analysis," *Electronics*, vol. 14, no. 11, p. 2284, 2025.

[20] W. C. Ouédraogo, K. Kaboré, Y. Li, H. Tian, A. Koyuncu, J. Klein, D. Lo, and T. F. Bissyandé, "Large-scale, independent and comprehensive study of the power of llms for test case generation," *Empirical Software Engineering*, 2026.

[21] Z. Xue, L. Li, S. Tian, X. Chen, P. Li, L. Chen, T. Jiang, and M. Zhang, "Llm4fin: Fully automating llm-powered test case generation for fintech software acceptance testing," in *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2024)*, 2024.

UNIVERSITAS GADJAH MADA
Rancang Bangun Sistem Pembangkit Skenario Pengujian Multi-Persona Berbasis Large Language Model (LLM)
Tsania Qurrota A'yunin Qulub, Divi Galih Prasetyo Putri, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

[22] Anthropic, "Reducing latency — claude platform docs," https://platform.claude.com/docs/en/test-and-evaluate/strengthen-guardrails/reduce-latency, 2026.

[23] L. Zheng, W.-L. Chiang, Y. Sheng *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023.

[24] N. B. Hasan, M. A. Islam, J. Y. Khan, S. Senjik, and A. Iqbal, "Automatic high-level test case generation using large language models," in *Proceedings of the IEEE/ACM 22nd International Conference on Mining Software Repositories (MSR 2025)*, 2025.

[25] A. Pereira, B. Lima, and J. P. Faria, "Apitestgenie: Generating web api tests from requirements and api specifications with llms," in *Proceedings of the 7th ACM/IEEE International Conference on Automation of Software Test*, 2026, pp. 46–54.

[26] D. Clerissi, E. Masserini, D. Micucci, and L. Mariani, "Automated testing of task-based chatbots: How far are we?" *arXiv*, 2026.

[27] S. Park *et al.*, "Fund2persona: A framework for building and refining financial advisor personas from fund disclosure data," *arXiv*, 2026.

[28] J. Gonnermann-Müller, J. Haase, N. Leins, T. Kosch, and S. Pokutta, "Maintaining stable personas? examining temporal stability in llm-based human simulation," in *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026.

[29] Z. Nan, Z. Guo, K. Liu, and X. Xia, "Test intention guided llm-based unit test generation," in *Proceedings of the IEEE/ACM 47th International Conference on Software Engineering (ICSE 2025)*, 2025.

[30] G. Ryan *et al.*, "Code-aware prompting: A study of coverage-guided test generation in regression setting using llm," *Proceedings of the ACM on Software Engineering (FSE)*, 2024.

[31] J. A. Pizzorno and E. D. Berger, "Coverup: Effective high coverage test generation for python," in *Proceedings of the ACM International Conference on the Foundations of Software Engineering (FSE 2025)*, 2025.

[32] A. F. Diansyah, M. R. Rahman, R. Handayani, D. D. N. Cahyo, and E. Utami, "Comparative analysis of software development lifecycle methods in software

UNIVERSITAS GADJAH MADA
Rancang Bangun Sistem Pembangkit Skenario Pengujian Multi-Persona Berbasis Large Language Model (LLM)
Tsania Qurrota A'yunin Qulub, Divi Galih Prasetyo Putri, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

development: A systematic literature review," *International Journal of Advances in Data and Information Systems*, 2023.

[33] M. O. Ahmad, D. Dennehy, K. Conboy, and M. Oivo, "Kanban in software engineering: A systematic mapping study," *Journal of Systems and Software*, vol. 137, pp. 96–113, 2018.

[34] Plaky, "Plaky: A centralized project management tool," https://plaky.com/, 2026.

[35] F. Dalpiaz and S. Brinkkemper, "Agile requirements engineering with user stories," in *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 2018.

[36] A. M. Maatuk and E. A. Abdelnabi, "Generating uml use case and activity diagrams using nlp techniques and heuristics rules," in *Proceedings of the 4th International Conference on Data Science, E-Learning and Information Systems (DATA 2021)*, 2021.

[37] Atagong *et al.*, "A review on knowledge and information extraction from pdf documents," *Frontiers in Artificial Intelligence*, 2025.

[38] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 1877–1901.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017, pp. 5998–6008.

[40] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016, pp. 1715–1725.

[41] Z. Wang, K. Ramnath, B. Bi, S. K. Pentyala, S. Chaudhuri, S. Mehrotra, Z. Zhu, X.-B. Mao, S. Asur, and N. Cheng, "Reinforcement learning for llm post-training: A survey," *arXiv preprint arXiv:2407.16216*, 2024.

[42] Anthropic, "Claude models overview," https://platform.claude.com/docs/en/models/overview, 2026.

[43] ——, "Claude haiku 4.5," https://platform.claude.com/docs/en/models/haiku-4-5/overview, 2026.

UNIVERSITAS GADJAH MADA
Rancang Bangun Sistem Pembangkit Skenario Pengujian Multi-Persona Berbasis Large Language Model (LLM)
Tsania Qurrota A'yunin Qulub, Divi Galih Prasetyo Putri, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

[44] ——, "Claude sonnet 4.6," https://platform.claude.com/docs/en/models/sonnet-4-6/overview, 2026.

[45] ——, "Claude opus 4.8," https://platform.claude.com/docs/en/models/opus-4-8/overview, 2026.

[46] ——, "Messages api — claude platform docs," https://platform.claude.com/docs/en/api/messages, 2026.

[47] ——, "Api overview — claude platform docs," https://platform.claude.com/docs/en/api/overview, 2026.

[48] ——, "Rate limits — claude platform docs," https://platform.claude.com/docs/en/api/rate-limits, 2026.

[49] J. White *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[50] Anthropic, "Claude prompting best practices," https://platform.claude.com/docs/en/build-with-claude/prompt-engineering/claude-prompting-best-practices, 2026.

[51] A. Neumann, Y. Pi, and J. Singh, "Who controls the conversation? user perspectives on generative ai (llm) system prompts," in *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026.

[52] W. Zhou, S. Zhang, H. Poon, and M. Chen, "Context-faithful prompting for large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[53] S. Das, S. Saha, and R. K. Srihari, "Diving deep into modes of fact hallucinations in dialogue systems," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022.

[54] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, "Self-refine: Iterative refinement with self-feedback," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

[55] W. Zou, Y. Liu, Y. Wang, Y. Chen, N. Gong, and J. Jia, "Pishield: Detecting prompt injection attacks via intrinsic llm features," *arXiv*, 2025.

UNIVERSITAS GADJAH MADA
Rancang Bangun Sistem Pembangkit Skenario Pengujian Multi-Persona Berbasis Large Language Model (LLM)
Tsania Qurrota A'yunin Qulub, Divi Galih Prasetyo Putri, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

[56] Anthropic, "Token counting — claude platform docs," https://platform.claude.com/docs/en/build-with-claude/token-counting, 2026.

[57] ——, "Pricing — claude platform docs," https://platform.claude.com/docs/en/about-claude/pricing, 2026.

[58] P. Verga, S. Hofstatter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis, "Replacing judges with juries: Evaluating llm generations with a panel of diverse models," *arXiv*, 2024.

[59] I. Otaduy and O. Díaz, "User acceptance testing for agile-developed web-based applications: Empowering customers through wikis and mind maps," *Journal of Systems and Software*, vol. 133, pp. 195–211, 2017.

[60] M. H. Thabibi, S. F. A. Wati, and T. P. Rinjeni, "Implementasi user acceptance testing (uat) pada website e-commerce umkm bbhealthy," *ATASI: Adopsi Teknologi dan Sistem Informasi*, vol. 4, no. 1, pp. 19–26, 2025.

[61] J. T. Almonte, S. A. Boominathan, and N. Nascimento, "Automated non-functional requirements generation in software engineering with large language models: A comparative study," *arXiv*, 2025.

[62] Z. Chen, C. Wang, W. Sun, X. Liu, J. M. Zhang, and Y. Liu, "Promptware engineering: Software engineering for prompt-enabled systems," *arXiv*, 2026.

[63] A. B. Belle, G. El Boussaidi, T. C. Lethbridge, S. Kpodjedo, H. Mili, and A. Paz, "Systematically reviewing the layered architectural pattern principles and their use to reconstruct software architectures," *arXiv*, 2021.

[64] Z. Gao, C. Bird, and E. T. Barr, "To type or not to type: Quantifying detectable bugs in javascript," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, 2017, pp. 758–769.

[65] M. I. Habib, A. A. Maruf, and M. J. A. Nabil, "An exploration into web session security: A systematic literature review," *arXiv*, 2023.

[66] Y.-Y. Liu, Z. Zheng, F. Zhang, J.-C. Feng, Y.-Y. Fu, J.-D. Zhai, B.-S. He, X. Zhang, and X.-Y. Du, "A comprehensive taxonomy of prompt engineering techniques for large language models," *Frontiers of Computer Science*, 2026.

[67] F. Lau, "Same input, different scores: A multi model study on the inconsistency of llm judge," *arXiv*, 2026.

UNIVERSITAS
GADJAH MADA
Rancang Bangun Sistem Pembangkit Skenario Pengujian Multi-Persona Berbasis Large Language Model (LLM)
Tsania Qurrota A'yunin Qulub, Divi Galih Prasetyo Putri, S.Kom., M.Kom., Ph.D
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

[68] A. K. Tiwari, "When consistency does not mean reliability: Evaluating local llm judges against human ratings," *arXiv*, 2026.

[69] G. Kohli, "Nine judges, two effective votes: Correlated errors undermine llm evaluation panels," *arXiv*, 2026.

[70] Anthropic, "Pricing," https://claude.com/pricing, 2026.

[71] A. Kapetanovic, K. Altwlkany, A. Mercep, T. Duricic, and E. Lacic, "Anchoring bias in llm-as-a-judge systems: Prior scores compromise evaluation independence," *arXiv*, 2026.

[72] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: Nlg evaluation using gpt-4 with better human alignment," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.