

A concept to simulate stakeholders in product development with personas and generative AI

Stefan Wellsandt^{1*}, Karl Hribernik¹, Jan Naumann², Klaus-Dieter Thoben¹

¹ BIBA - Bremer Institut für Produktion und Logistik GmbH at the University of Bremen, Bremen, Germany

² Institut Technik und Bildung (ITB), University of Bremen, Bremen, Germany

*Corresponding author. E-Mail address: wel@biba.uni-bremen.de

Abstract: The article proposes simulating stakeholder workshops in product development using persona-based generative-AI agents. It argues that real workshops are hard to coordinate and that early-stage concept creation could benefit from a scalable, repeatable alternative. The approach instantiates multiple persona-driven agents representing distinct stakeholder roles (e.g., user, designer, safety engineer, and cost analyst) within an agent-based simulation framework (TinyTroupe). Agents are guided through a structured workshop workflow: setup, document review, idea generation, detailed recommendations, trade-off reasoning, and prioritization. During the process, agents can ask targeted questions to a human facilitator, and the answers can be stored in a shared knowledge base to improve consistency across rounds. A demonstration on an educational SCARA robot kit shows how to produce ranked improvement recommendations and synthesize different stakeholder perspectives into a consolidated outcome. The paper generalizes the workflow into a concept model and identifies key research gaps: generating empirically grounded personas, measuring how persona traits influence outcomes, maintaining persona consistency over long dialogues, validating outputs against human baselines, and scaling to richer product representations and tool access (e.g., structured product models and engineering artifacts). It also notes limited generalizability because the evaluation covers only one scenario and lacks systematic quality metrics.

Keywords:

Generative AI, Personas, Stakeholder Simulation, Product Development, Multi-Agent Systems

1 Introduction

Dialogue among stakeholders of complex products is essential for anticipating, reviewing, and resolving problems. Involving stakeholders from the product lifecycle during product development activities is costly and time-consuming. The involved persons need to invest their work time (employees) or personal time (users) and might need to travel to a physical location to work on a specific task, for instance, during a workshop or test session. A second issue is that persons, especially those with specific expertise or rare characteristics, e.g., senior experts or members of a specific user group, may not be available for an activity, which may cause delays or reduce the quality of the activity's results.

One approach to address these issues is to simulate stakeholders [1,2]. Product development already applies simulation techniques such as roleplaying with personas representing typical individuals within a target group. If humans adopt a role like this, the quality of the results might improve [3,4], but someone must still invest their time at a cost.

Recent research indicates that generative AI can create text to describe stakeholder personas and human behavior [5,6]. Also, there are libraries for simulating interactions between generative AI personas (e.g., TinyTroupe, Concordia, and AgentSociety), but their applicability in product development remains

unclear. If computer-based stakeholder simulation in product development is feasible, affordable, and on par regarding results quality, it may have a substantial positive impact on costs and time, due to its scalability.

The **goal** of this article is to present a concept for simulating stakeholders in a product development environment with multiple generative AI personas. Our guiding research questions are as follows. Is the simulation of a realistic product development activity technically feasible? What are the key characteristics of related simulation experiments? Which open issues require deeper investigations?

Section 2 outlines the research problem, and Section 3 introduces the chosen methods and approaches. The results in section 4 cover the technical setup and present important findings from the experiments, which are generalized in a concept illustration. Section 5 discusses key aspects with respect to this article's goals and open issues. This article concludes by acknowledging the limitations and providing an outlook.

2 Research Problem

Computer-based simulation of stakeholders in product development is not yet common, largely due to the assumed effort required to train domain-specific probabilistic and rule-based AI models. Crowder et al. [1] built a complex agent-based simulation for an engineering team that can exchange abstract information to perform tasks and sub-tasks. Liu et al. [2] present a simulation for user engagement in new product development with a high level of abstraction. Neither simulation covers how simulated stakeholders actually contribute to an activity in terms of content.

Using pre-trained (generative) AI – such as ChatGPT, Claude, or DeepSeek – to simulate the concrete contributions of a stakeholder can help address part of the problem outlined above. These software systems can generate convincing content from user prompts. Prompts may contain task descriptions, product outlines, persona profiles, and other useful information to consider in responses. A key challenge for using pre-trained models effectively for this purpose and domain is that many activities involve multiple stakeholders concurrently. Example activities are brainstorming, design review, and root-cause analysis. This observation means that a simulation concept must also handle the interactions among simulated stakeholders. There are simulation libraries that claim to be suitable; for instance, Microsoft's TinyTroupe project indicates it can simulate focus groups to assess product designs [7]. The academic literature offers very limited insight into the applicability of computer-based multi-stakeholder simulation in product development.

3 Methods and Approaches

This research is grounded in multi-agent frameworks supporting generative AI, such as TinyTroupe, which provide basic functions for managing persona-driven agents, message exchange via conversations, and data exchange with one or more language models. In this context, agents are software with limited autonomy that can execute agent tools to interact with other software. These tools are software functions with semantic descriptions that may, for instance, interact with a virtual product model and return information for subsequent reasoning activities. We use a scenario to narrow our research scope on a specific development activity, work environment, and product.

3.1 Scenario

Activity. We chose a product improvement task rather than a new product development task to reduce the simulation's complexity. Improvement requires substantially less information and time; such tasks are also significantly more common in practice.

Environment. The starting point for our simulation is a typical situation encountered in product improvement. A group of experts needs to improve an existing product to better address user needs. The group involves users directly, as suggested by user-driven design practices. The interaction format is a workshop because it encourages the exchange and consolidation of understanding, opinions, needs, and constraints. Such workshops are acknowledged in product development and allow participants to interact, but typically do not enforce interactions.

The workshop has sequential phases, each with one specific objective to fulfill:

1. Intro: Share setup and scenario with participating agents
2. Document review: Agents review task-related documentation individually
3. Initial thoughts: Agents share initial improvement ideas
4. Specific recommendations: Agents identify detailed improvements
5. Cost-benefit analysis: Agents analyze cost, time, and manufacturability
6. Prioritization: Agents rank each improvement

These phases ensure the environment is rich and challenging, as they allow participants to prepare, gradually detail their findings, quantify them, and mutually agree on a final result.

Product. The example product is a custom-made small SCARA robot arm built from 3D-printed plastic parts and (standard) components, such as screws, nuts, bearings, tools, and electric drives. We used this robot in educational settings where students must assemble the robot parts with support from a voice-enabled, intelligent digital assistant. A 7-page manual is available that contains assembly instructions and didactic/explanatory information on the purpose and use of tools and components. Figure 1 illustrates the assembly with components and sub-assemblies.

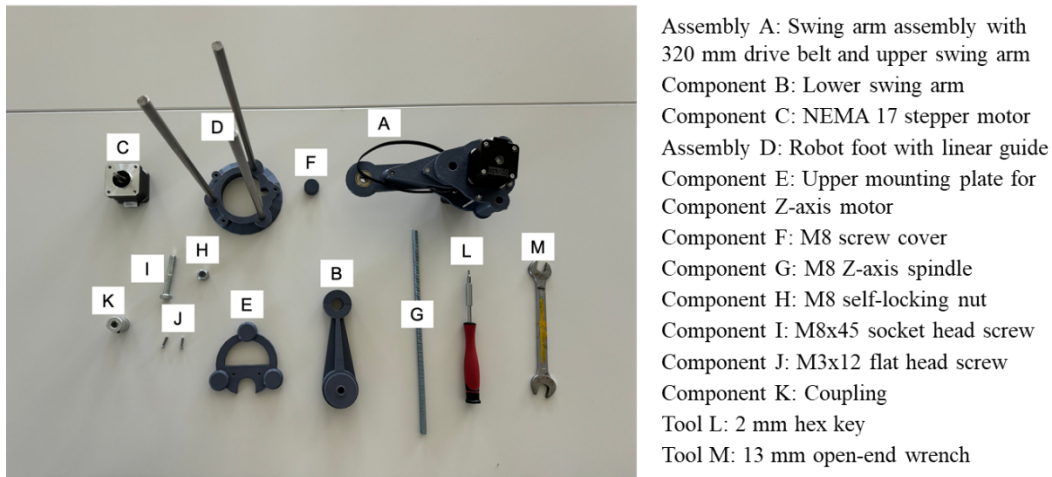


Figure 1: Educational table-top robot assembly used as a scenario for the experiments

3.2 Simulation

The simulation approach relies on using an existing agent-based simulation library that supports generative AI models. This library must provide basic facilities for persona management, agent creation from personas, and communication with the AI model via a data exchange interface. Besides, it must provide memory management across repeated interactions (but not experiment runs). The simulation environment models the workshop above with multiple participants and specific roles. Each phase has

a fixed number of simulation rounds, each of which provides an opportunity for agent reasoning and interaction with other agents or a human. The agents decide to interact based on their reasoning.

In the following sections, we use “GenAI system” to refer to a system that generates text responses from text input. It is not only a large language model (LLM) in the narrow sense, but a complex system of the LLM and undisclosed pre- and post-processing steps in the case of commercial providers like OpenAI.

4 Results

The concept’s basis is a series of iterative experiments with Microsoft’s experimental TinyTroupe simulation library for Python [8]. It leverages large language models (LLMs) and orchestrates agents with configurable traits. Agents can interact in environments called “TinyWorlds”, engage in conversations, receive broadcasts, and perform actions. The library enables product brainstorming, user testing simulations, market research, and the generation of training data. Figure 2 illustrates the technical experiment setup.

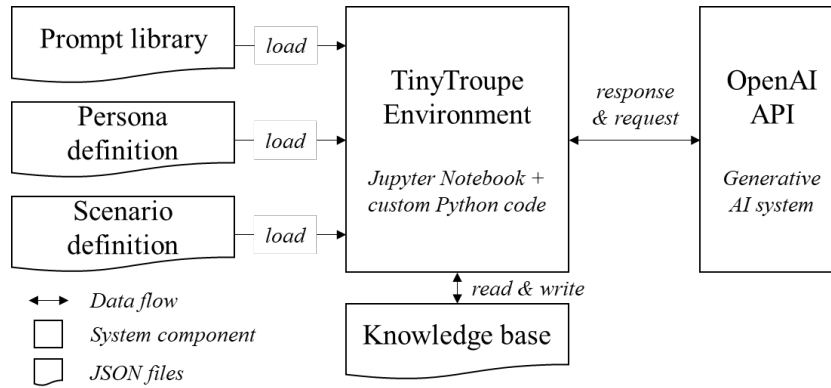


Figure 2: Technical setup for experimenting with personas and generative AI agents

We implemented two custom human-in-the-loop processes. First, the user receives a prompt to provide meaningful free-text input after each phase. This ensures a general intervention option for intermediate results. Second, if an agent’s response includes a question for the facilitator (simulation user), the user can provide an answer or skip it. Answered questions are broadcast and stored in a knowledge base that is available across experiments. A custom token tracker helps understand simulation costs per experiment and phase. We generated personas iteratively with the pretrained GenAI system Claude Opus 4.5 to avoid the substantial effort of hand-crafted, data-grounded persona design. The latter may yield more accurate personas, but since our goal was to create a sound concept first, we opted for AI-supported generation. The prompting was iterative, i.e., we refined the prompts based on intermediate persona results. Our main goal for this activity was to design persona-generation prompts that result in a coherent set of stakeholders. We assessed the coherence based on experience.

Personas contain a broad range of attributes, including demographics, occupation, education, goals, personality traits, preferences, beliefs, and communication style. We experimented with a team of up to seven agents selected from the persona pool in Table 1. The background summary outlines each persona, but was not used in the experiments (instead, we used concrete attributes). We set the limit to seven agents to balance persona variety and expected token costs. Identifying the technical limit was not a goal

Table 1: Persona pool

No.	Role	Age	Gender	Background Summary
1	Vocational Training Student - Mechatronics	19	Male	First-year vocational training student (Auszubildender) with basic knowledge of hand tools and simple electronics. This is his first exposure to robotics, and he’s excited but nervous about making mistakes.

2	Engineering Student - Robotics	23	Female	Fourth-year robotics engineering student with hands-on experience, including building several robot projects and a robot arm for a competition. Completed an internship at an industrial automation company.
3	Professor of Engineering Education	48	Male	Professor with 20 years of experience teaching robotics and automation. Designs curricula for vocational training programs and published research on hands-on learning effectiveness.*
4	Educational Kit Designer	36	Female	Designer with a Master's in Industrial Design, specializing in design for education. Designs educational robotics kits used by students across Europe, focusing on making technology accessible through thoughtful design.
5	Safety and Quality Engineer	45	Male	Certified safety engineer with 20 years evaluating products for compliance with EN 71, machinery directives, and educational laboratory standards. Dedicated to preventing harm through proper design and risk assessment.*
6	Curriculum Developer - STEM Education	38	Female	Curriculum developer specializing in robotics and automation training. Created hands-on learning modules used in vocational schools and universities across Germany.
7	Procurement and Cost Analyst	41	Male	Procurement and cost analyst for educational equipment. Optimizes budgets for STEM education programs across multiple institutions, focusing on maximizing educational value per euro spent.
* Note the unrealistic attributes for age (48, 45) with the "20 years of experience" occupation description. This is not a typo, but an actual result of the persona generation process we did not realize during the experiment phase.				

We ran all experiments using OpenAI's gpt-4.1-mini model to limit token costs and response generation time while maintaining reasonable response quality (we did not systematically investigate response quality as outlined in Section 5). Team compositions are Minimal (#1, #3), Student Focus (#1, #2, #3), Design Review (#3, #4, #5), Curriculum Alignment (#2, #3, #6), Budget Optimization (#3, #4, #7), and Full (#1 to #7). We did not empty the knowledge base between experiments to test knowledge accumulation. Token limits for input (e.g., conversation history and instructions) and output (response length) remained the same. Each run concludes with a ranked list of recommendations, each with a textual description of the suggested improvement and a statement and explanation of the expected impacts on time, cost, and feasibility. Table 2 summarizes the experiment runs.

Table 2: Experiment run examples

No.	Team	Rounds	Recommendations	Token Count	Costs USD	Duration
1	Minimal	11	8	796.826	0,35	00:10:02
2	Student Focus	16	6	907.913	0,39	00:13:03
3	Design Review	15	6	977.613	0,42	00:11:07
4	Design Review	22	7	949.899	0,41	00:08:27
5	Curriculum Alignment	16	4	988.671	0,43	00:10:12
6	Budget Optimization	18	2	920.025	0,40	00:08:53
7	Budget Optimization	11	9	923.078	0,40	00:09:40
8	Full	17	19	1.023.150	0,44	00:10:35
9	Full	22	9	1.126.857	0,49	00:11:56
10	Student Focus	11	3	1.677.528	0,73	00:20:02

Figure 3 presents contributions from the Document Review phase to outline the specificity and length of the agents' contributions. This example includes THINK and TALK messages (there are several others in TinyTroupe). The former is an agent's response to another agent's input, and the latter is the message meant for another agent. Agent contributions aim to meet the workshop instructions, for instance, focusing on three specific components or processes from the assembly manual.

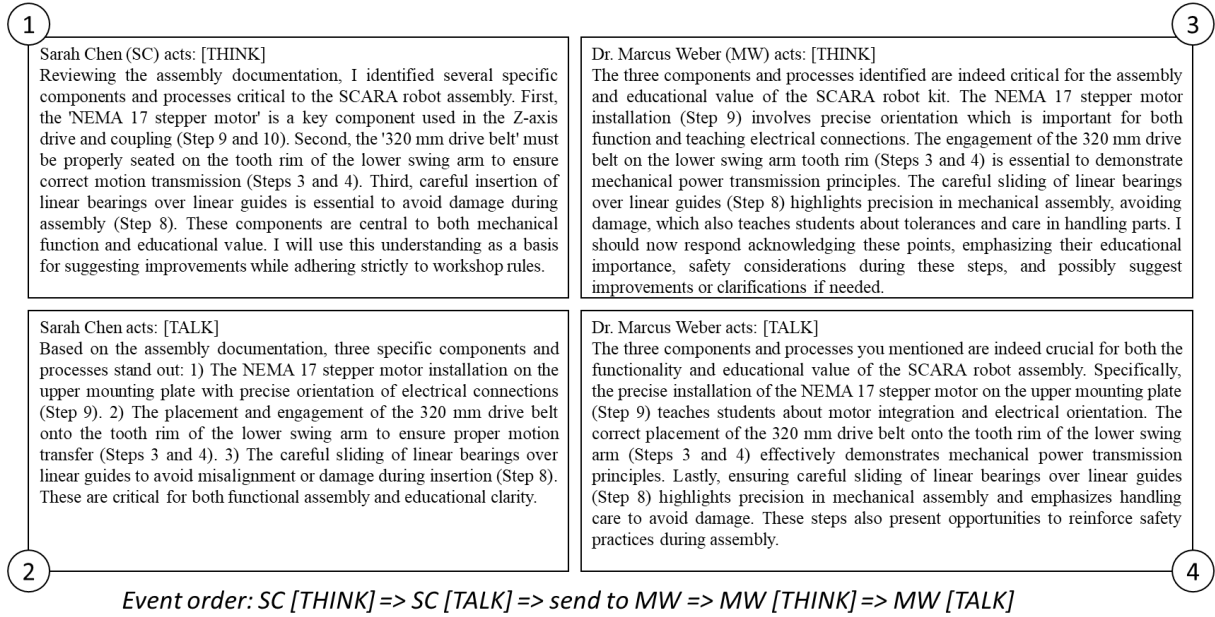


Figure 3: Example contributions between Designer [Sarah Chen] and Educational Expert [Marcus Weber]

Figure 4 illustrates key parts of the simulation flow for experiment run #1 with summarized example contributions. The horizontal arrows indicate that agent contributions are passed from agent to agent, and subsequent responses build on a previous one. Vertical ones represent sequential phase execution.

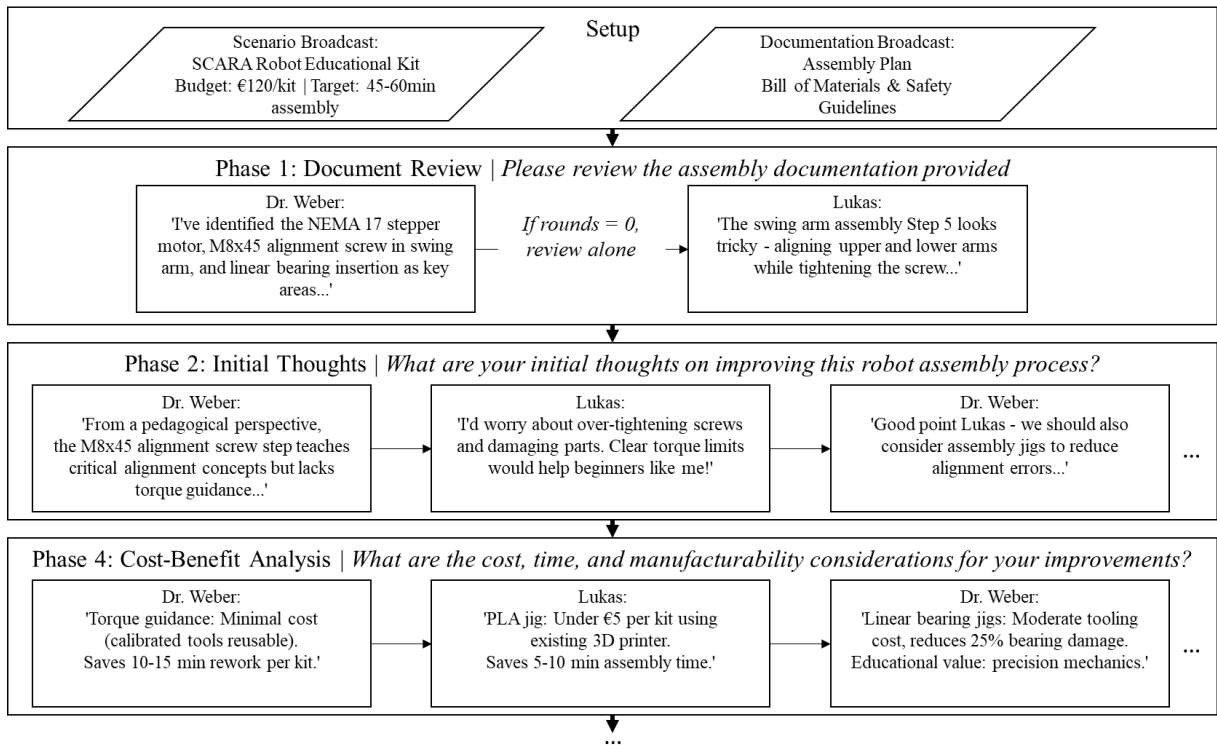


Figure 4: Conversation flow and summarized example content from an experiment with the Minimal team

Since the illustration above is specific to our setup, Figure 5 aims to generalize all experiment findings into a concept. Each phase relates to one objective that the agents aim to fulfil with conversations extending over one or more rounds. The last phase integrates previous results to generate a single final

result, e.g., a ranked list of recommendations. A human facilitator may answer questions upon an agent's request or provide input at the end of each phase.

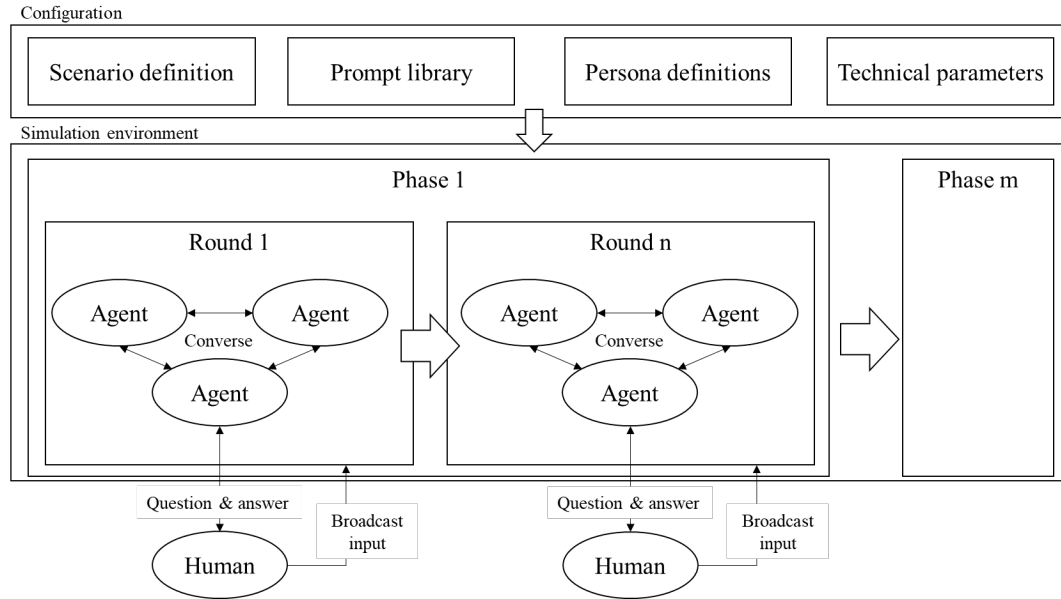


Figure 5: Concept for simulating stakeholders with persona-based agents and generative AI

5 Discussion

This section outlines discussion topics that emerged from the experiments. These topics cover technical aspects, design decisions for personas and scenarios, result quality, and scalability.

5.1 Technical setup

TinyTroupe is a good starting point for feasibility tests and pre-studies for the concept creation. However, the library hasn't received updates from Microsoft for several months, while other agent-based simulations are actively being developed. Switching to another simulation library might provide more functions and control over experiments. A second aspect of the setup is that it lacks a dedicated product model, which could have a substantial positive impact on the simulation by organizing product information for agents and allowing access to specific engineering information. Integrating it depends on agent tools for accessing and managing the product model, for instance, to read a bill of materials. Complex information like product geometry and physics-related simulation is already being integrated into generative-AI systems via multimodal Retrieval Augmented Generation (RAG) and tool-augmented CAD interfaces [9,10]. Its readiness and compatibility with agent-based simulation are unclear.

5.2 Persona design and behavior

Data-grounded personas. An accurate persona design should be grounded in acknowledged data about the target stakeholders to reduce bias and minimize assumptions [11]. Examples include gender and age distributions, country of residence, and typical educational background. The downside is that data must be available and accessible to a human to hand-craft the persona traits, or to a software tool (e.g., a GenAI system) to automatically generate a persona. Another aspect is that the more features the persona has, the more relevant it is to check if features are dependent. If they are, generating a consistent persona would benefit from grounding in data or at least specific reasoning or plausibility checks. For example, a persona's age and experience: a young persona with extensive experience might seem unrealistic in many scenarios. Our experiments yielded no clear results [12] on whether the GenAI system used an internal representation of these dependencies (e.g., in its training data) when generating the personas.

However, Table 1 includes two controversial personas, where the age and experience stated in the occupation description might raise plausibility questions (assuming that becoming a professor at 28 years is unlikely but not impossible).

Targeted persona design. Changes to persona traits have unknown consequences for the simulation and its results. It is not yet clear whether the experiment designer can estimate the impact of persona changes before running the experiment. Investigating this would likely require a deeper analysis of how each specific GenAI model represents personas and their conversations (e.g., [13]). Such a research activity is highly ambitious and would require tools to inspect a model's internal working, which is unlikely for commercial models from, for instance, OpenAI or Anthropic, and more likely for models with open weights, like Qwen (Alibaba Cloud) or DeepSeek (DeepSeek Ltd.). On the positive side, the uncertainty above may lead to interesting and highly desirable emergent situations that typically do not occur in scenarios with humans, for instance, because their traits do not provide opportunities for such emergent phenomena. Also, systematically running many simulations with edge-case traits may lead to trait combinations that drive the conversation in entirely unexpected directions, revealing something that accessible human participants could never conclude. Such results would be relevant when the number of target stakeholders is high and diverse, while the pool of accessible human participants is small (e.g., due to budget constraints or missing contact).

Persona integrity for agents. GenAI systems used in persona-based role-playing, such as the simulated workshop above, may respond in ways inconsistent with their persona [14]. This integrity was not specifically investigated in the experiments, but it is a critical aspect of an authentic simulation. The impact of inconsistencies on simulation results is unclear and requires further investigation, potentially using new tools to examine GenAI system behavior.

5.3 Experiment design

Choosing simulated participants for an experiment is a key decision in the simulation. While the team composition among the already-defined personas depends on the task and scenario, different teams are likely to produce better or worse simulation results. As with the persona design and behavior, it is not yet clear whether the impact of team composition on results can be reliably anticipated.

5.4 Result quality

We did not focus on the quality of the simulation results in this study, especially compared to human stakeholders. Our observations and results indicate that the generated recommendations are concrete and appear actionable (we did not implement them to verify). Further analysis of result quality could leverage the generated explanations for each expected impact and seek human expert opinion (e.g., acceptable vs. unacceptable). Another option is for humans to create a validation dataset (gold standard dataset) in which they, for instance, provide their own recommendations and explain the expected impacts based on their expertise. This approach would require a comparative study where a human team competes with a simulated team – a hybrid team could also provide useful insights, but requires a more complex technical setup where agents can monitor human conversations and intervene.

Another aspect of result quality is that the generated recommendations are not the result of a logic-driven process but rather a sequential refinement of a single final prompt sent to the GenAI system to generate tokens representing recommendations. This is an inherent characteristic of the selected simulation library.

Finally, quantitative values in the expected impact (e.g., 15% or 5 minutes less) are hard to verify because the rationale is missing from the agent responses, and the experiment input intentionally does not provide baselines. The agents sometimes asked the facilitator for such details, though. We observed

this often when the agents were not strictly encouraged to rely on their role expertise – they asked for details at almost every step, and we decided to modify the system prompt to avoid this.

5.5 Scalability

Technical factors: The simulation’s ability to work effectively across a wide range of development workshop situations is a requirement for successful application and adoption in practice. We identified two drivers of scalability: simulation time and costs. The concept has several parameters that influence these drivers, such as the number of participating agents, the number of simulation rounds, the selected GenAI system, and input and output token limits. Typically, more agents and rounds result in more API calls to the model, driving the token count and related costs. Our results in Table 2 do not clearly substantiate this pattern, especially because the experiment runs #1, #7, and #10 differ significantly, even though they each have 11 rounds. We did not investigate these data points further. Another typical assumption is that a GenAI system that generates responses of potentially higher quality typically takes longer and costs more. Higher token limits for input and output increase time and costs; however, they also influence the quality of the results, including reasoning quality and response length.

Non-technical factors: In addition to the technical parameters above, preparing accurate, data-grounded personas will likely affect the time and cost of the simulation. This effort is probably a one-time investment that would pay off the more simulations reused personas. There are already initiatives to create persona databases to address this issue [15]. Also, the impact of assessing AI-based simulation results (e.g., validation, review, and interpretation) may not scale if intense human involvement is desired due to a lack of trust in the results or because it is required by law, compliance, or contracts. Software tools for automated result assessments may increase the scalability to some extent, though. Examples include validation against existing knowledge bases, standards, guidelines, or trusted deterministic simulations.

6 Conclusion

The experiments and the related concept provide a first systematic investigation for using generative AI and personas to simulate stakeholders in product development. While the initial results are a good starting point, several limitations must be considered to better understand the merits of this concept. *First*, this study covered only one scenario and one product, limiting the generalizability. Also, it doesn’t cover product-services or product-related software, which are essential components of many complex value offers and smart products. *Second*, the simulation does not allow iterations across phases, i.e., agents or the human facilitator cannot decide to repeat a phase and apply more refined information. *Third*, the experiments could be more systematic in their use of the knowledge base, e.g., by cleaning the database between steps to investigate how agents ask the facilitator questions. *Fourth*, there was no systematic assessment of result quality because we had not yet identified adequate metrics and tools. *Fifth*, the simulation lacks integrations with a formal product model concerning function, structure, and behavior. Future integrations may cover other computer-based simulations along the product life cycle, ranging from engineering simulations to simulated product usage. Addressing the above limitations could lead to further automation of product development and shorter product lifecycles.

Acknowledgements

This work is funded by the European Union's Horizon Europe project WASABI “White-label shop for digital intelligent assistance and human-AI collaboration in manufacturing” (Grant agreement No 101092176). The work presented here reflects only the authors’ view and the European Commission is not responsible for any use that may be made of the information it contains.

References

- [1] R. M. Crowder, M. A. Robinson, H. P. N. Hughes and Y.-W. Sim, "The Development of an Agent-Based Modeling Framework for Simulating Engineering Team Work," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 42, no. 6, pp. 1425-1439, 2012. DOI: <https://doi.org/10.1109/TSMCA.2012.2199304>.
- [2] Y. Liu, B. S. Onggo and J. Busby, "Agent-based modelling of user engagement in new product development," *Technovation*, vol. 135, p. 103062, 2024. DOI: <https://doi.org/10.1016/j.technovation.2024.103062>.
- [3] M. Buchenau and J. F. Suri, "Experience prototyping," in *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*: New York, NY, USA: ACM, 2000, pp. 424-433.
- [4] J. Pruitt and J. Grudin, "Personas," in *Proceedings of the 2003 conference on Designing for user experiences*: New York, NY, USA: ACM, 2003, pp. 1-15.
- [5] A. Schmidt, P. Elagroudy, F. Draxler, F. Kreuter and R. Welsch, "Simulating the Human in HCD with ChatGPT: Redesigning Interaction Design with AI," *Interactions*, vol. 31, no. 1, pp. 24-31, 2024. DOI: <https://doi.org/10.1145/3637436>.
- [6] J. Shin, M. A. Hedderich, B. J. Rey, A. Lucero and A. Oulasvirta, "Understanding Human-AI Workflows for Generating Personas," in *Designing Interactive Systems Conference*: New York, NY, USA: ACM, 2024, pp. 757-781.
- [7] P. Salem, R. Sim, C. Olsen, P. Saxena, R. Barcelos and Y. Ding, "TinyTroupe GitHub Project. LLM-powered multiagent persona simulation for imagination enhancement and business insights,," [Online]. Available: <https://github.com/microsoft/TinyTroupe>. [Accessed 22 January 2026].
- [8] P. Salem, R. Sim, C. Olsen, P. Saxena, R. Barcelos and Y. Ding, *TinyTroupe: An LLM-powered Multiagent Persona Simulation Toolkit*, 2025.
- [9] C. Kienle, B. Alt, D. Katic, R. Jäkel and J. Peters, *QueryCAD: Grounded Question Answering for CAD Models*, 2024.
- [10] S. Li and J. Corney, "MechRAG: a multimodal large language model for mechanical engineering," *Communications engineering*, vol. 4, no. 1, p. 187, 2025. DOI: <https://doi.org/10.1038/s44172-025-00517-z>.
- [11] B. J. Jansen, J. Salminen, S. Jung and K. Guan, *Data-Driven Personas*, 1st ed. 2021Cham: Springer International Publishing; Imprint: Springer, 2021.
- [12] A. S. Vezhnevets, J. P. Agapiou, A. Aharon, R. Ziv, J. Matyas, E. A. Duéñez-Guzmán, W. A. Cunningham, S. Osindero, D. Karmon and J. Z. Leibo, *Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia*, 2023.
- [13] S. Lee, S. Lim, S. Han, G. Oh, H. Chae, J. Chung, M. Kim, B. Kwak, Y. Lee, D. Lee, J. Yeo and Y. Yu, *Do LLMs Have Distinct and Consistent Personality? TRAIT: Personality Testset designed for LLMs with Psychometrics*, 2024.
- [14] P. Bhandari, N. Fay, M. Wise, A. Datta, S. Meek, U. Naseem and M. Nasim, *Can LLM Agents Maintain a Persona in Discourse?*, 2025.
- [15] J. Rupprecht, L. Fröhling, C. Wagner and M. Strohmaier, *German General Personas: A Survey-Derived Persona Prompt Collection for Population-Aligned LLM Studies*, 2025.