Human-Computer Interaction & Emerging Technologies, Vol. 233, 2026, 21–30
https://doi.org/10.54941/ahfe1007500

Context-Aware LLMs for Healthcare Requirements Engineering

Valeria Resendez[1], Andrew Hornback[2], Harinishree Sathu[2], J. Ben Tamo[2], Yining Yuan[2], May Wang[2] , Nese Baz[1], Funda Yildirim[1], Russell Chan[1], Maria Fernanda Cabrera[3], and Simone Borsci[1,4]

[1]University of Twente, Enschede, Overijssel, NL
[2]Georgia Institute of Technology, Atlanta, GA, USA
[3]Universidad Politécnica de Madrid, Madrid, Spain
[4]Imperial College London, London, UK

ABSTRACT

Requirements engineering (RE) is a collaborative, context-dependent, and resource-intensive process, particularly in highly regulated domains such as healthcare. Recent advances in large language models (LLMs) have raised questions about their potential in supporting early-stage requirements elicitation. However, integrating LLMs introduces an additional mediation layer between contextual knowledge and articulated system requirements. Drawing on Norman's concepts of the gulf of execution and the gulf of evaluation, this study examines under what contextual conditions LLMs approximate human expert–elicited requirements. We conducted a 3 × 3 × 3 simulation study comparing three LLMs (GPT-5.2, Claude 4.5 Sonnet, and Gemini 3 Pro), three knowledge conditions (none, proposal-based, and literature-based), and three expert-role prompts (none, pediatrician, and geneticist). Each combination was repeated 50 times, producing a total of 1,350 outputs. Results show significant variation in requirement quantity across models and knowledge conditions, but consistently low semantic alignment with human expert requirements. Retrieval-augmented knowledge reduced output volume without improving the alignment with human-expert requirements. Role prompting produced marginal effects. All models demonstrated high within-condition reliability, indicating stable but moderately aligned outputs. These findings suggest that LLMs could function more as tools to generate requirements for scaffolding than as expert emulators. While LLMs do not operationalize contextual knowledge into expert-level requirements, they may support early RE processes.

Keywords: Requirements elicitation, Human-AI collaboration, Large language models, Retrieval-augmented generation, Precision medicine, Stakeholder requirements

INTRODUCTION

Requirements engineering is a collaborative, context-dependent process that involves multiple stakeholders (Linåker et al., 2020; Sharp et al., 1999) and guides technology design by translating user needs into system requirements (Sommerville, 2016). This translation is resource-intensive and requires continuous interdisciplinary collaboration across experts such as paediatricians, developers, data managers, and product owners

Received February 21, 2026; Revised April 6, 2026; Accepted April 24, 2026; Available online July 20, 2026

(Karolita et al., 2024, Cutting et al., 2016; Danahey et al., 2017; Meyer et al., 2012). Challenges are amplified in highly regulated environment such as healthcare, where requirements must align with complex regulatory, technical, and clinical constraints, and early errors can have serious consequences (Van Velsen et al., 2013; Cysneiros, 2002).

Given these challenges, researchers are exploring Large Language Models (LLMs) to computationally support early-stage requirements engineering (Alhoshan et al., 2025; Arora et al., 2023; Quattrocchi et al., 2025). Emerging literature suggests that LLMs can generate, classify, and evaluate requirements, thereby complementing activities that have traditionally relied solely on human expertise (Alhoshan et al., 2025; Arora et al., 2023; Quattrocchi et al., 2025). Beyond text generation, LLMs can emulate human behavior, including demographic, social, and communication characteristics (Hu & Collier, 2024; Mouri Zadeh Khaki et al., 2025), suggesting potential roles in simulating stakeholder input.

However, integrating LLMs into requirements engineering introduces an additional mediation layer between stakeholder needs and the resulting requirements. In early phases of requirements engineering, experts derive requirements by analyzing textual information such as scientific literature, reviews, project proposals, or product documentation (Carley & Palmquist, 1992). Through this analysis, the contextual knowledge is articulated into user needs. At this stage, LLMs could be integrated into the requirements engineering process by a retrieval-augmented generation (RAG) approach, where relevant documents are retrieved and embedded into the LLMs' context (Abo El-Enen et al., 2025). This way, LLMs could function as an alternative mechanism for translating knowledge into user requirements. Yet, using LLMs in this way positions them as intermediaries, mediating between the contextual knowledge and an initial set of requirements (Chen et al., 2025).

From a human–computer interaction perspective, the intermediary role of the LLMs could be evaluated through Norman's concepts of the gulf of execution and the gulf of evaluation. The gulf of execution refers to the gap between users' intentions and the actions available to realize them, while the gulf of evaluation concerns the gap between a system's output and users' ability to interpret and assess it (Norman, 1992; Subramonyam et al., 2024). In LLM-supported requirements, if the model fails to adequately operationalize contextual knowledge into quality requirements, the execution gap widens. Similarly, if the generated requirements lack clarity, structure, or domain alignment, the evaluation gap increases. In both cases, the interaction cost of using LLMs rises, potentially offsetting the anticipated efficiency gains (Subramonyam et al., 2024).

Moreover, if LLMs introduce an additional mediation layer, the interaction costs they entail must be justified by the quality of the output generated. It is therefore necessary to examine how effectively LLMs execute the task of requirement generation. Specifically, we ask under what contextual conditions do LLMs produce requirements that approximate those elicited by human experts? To evaluate the LLM layer, we operationalize the gulf of execution as the model's ability to translate contextual knowledge into requirements that semantically align with expert-generated requirements. We

operationalize the gulf of evaluation as the consistency and interpretability of outputs across repeated iterations. We then manipulate contextual knowledge (none, proposal, literature) and expert role prompting to examine whether these reduce the execution gap.

METHODOLOGY

Through a 3 × 3 × 3 simulation study, we compared LLM-based requirements across different levels of contextual knowledge and simulated expert assumptions. Across three LLMs (e.g., GPT, Claude, Gemini), we tested three knowledge conditions: (a) no external knowledge, (b) summary project-proposal knowledge, and (c) domain-literature knowledge simulating three levels of expertise: (a) no explicit simulated expert, (b) paediatrician, and (c) geneticist. The 27 conditions (e.g., model–knowledge–expert combinations) were tested using the POE AI's API. Additionally, to capture variability in requirement generation, we repeated each condition combination 50 times. The iterations across conditions resulted in 1,350 outputs. Once the simulations were complete, we combined results across conditions and reported mean estimates alongside uncertainty measures (e.g., confidence intervals).

The models were selected for their complementary capabilities: GPT-5.2 excels in general-purpose, tool-based reasoning; Gemini 3 Pro offers multimodal integration and large-context reasoning; Claude Sonnet 4.5 emphasizes instruction fidelity and contextual sensitivity (Anthropic, 2025; Chartier et al., 2025; Sobo et al., 2025). Comparing these models allowed us to assess differences in LLM performance in requirements generation.

Simulated Experts

We created expert roles using an automated pipeline that translated a set of real-world experts into a single group of experts. Three steps were part of the process: (1) Role definition. We operationalize an expert group as a set of real-world experts who share a common disciplinary focus. (2) Information retrieval. We collected all publicly available information about selected experts, including professional biographies and relevant publications. (3) Synthesis. We synthesized the collected information into a description that reflected the group's collective expertise, priorities, and predefined focus. Using this process, we simulated two expert roles: pediatricians and geneticists. Pediatricians are physicians who specialize in the medical care of infants, children, and adolescents (Uchitel et al., 2022). Geneticists play a role in linking clinical and molecular insights (Castle et al., 2025).

Task and Evaluation

All models were prompted using a standardized template applied across conditions. The template specified the target system, the desired output format, and the expert role when applicable. Contextual knowledge was provided as a structured prefix. The prompt instructed the models to generate functional and non-functional requirements for a secure, privacy-preserving federated

infrastructure for health and genomic data across Europe. No restrictions were imposed on the length of individual requirements, allowing the models to determine statement granularity autonomously. Outputs were compared to a reference list of 366 requirements. These human-generated requirements were assembled over 18 months through literature review and expert input (~60 participants) collected via surveys, workshops, and focus groups.

Data Analysis

We evaluated the generated requirements using three outcome measures: requirement frequency, semantic alignment with expert requirements, and within-condition semantic consistency.

a) **Requirement frequency.** We examined whether providing contextual knowledge and assigning a simulated expert role influenced the number of generated requirements. To account for the count-based and condition-dependent variation in requirement frequency, we fitted a negative binomial regression model. The binomial model included the LLM type (GPT-5.2-instant, Claude-Sonnet-4.5, Gemini-3-Pro), knowledge condition (no knowledge, proposal, literature), and expert role (no expert, pediatrician, geneticist). We report results as Incidence Rate Ratios (IRRs), indicating how the expected number of generated requirements changes relative to the reference condition (Hilbe, 2011). Values above 1 indicate an increased generation rate, and values below 1 indicate reduced rates. For instance, an IRR of 1.20, for example, indicates a 20% increase, whereas an IRR of 0.80 indicates a 20% decrease.

b) **Semantic similarity to human expert requirements.** We assessed the semantic similarity between LLM-generated requirements and expert-elicited requirements. We operationalized similarity by computing the cosine similarity between each generated requirement per condition and all requirements in the expert reference set. Each generated requirement was evaluated against the best match from the human generated requirements. Requirements with similarity ≥ 0.70 were considered semantically aligned (Cer et al., 2017; Reimers & Gurevych, 2019). For each iteration, we computed the percentage of aligned requirements. We report the summaries per condition using mean and standard deviation.

c) **Similarity within-conditions.** We evaluated the semantic similarity of the LLM-generated requirements across iterations within the same condition. Each requirement was first converted into a sentence embedding using a pre-trained transformer model. For every iteration, embeddings were averaged to create a single representation capturing the semantic content of that run. This averaging step minimizes the impact of small wording differences while preserving the main semantic themes. As a next step, we computed a condition-level centroid by averaging the iteration-level embeddings across all repetitions. Semantic consistency was measured as the cosine similarity between each iteration's embedding and the corresponding condition centroid (Tehenan, 2025). Higher cosine similarity values indicate greater consistency with the typical content generated under that condition.

RESULTS

Requirement Frequency

A negative binomial regression was used to examine differences in the number of requirements generated across conditions. The model was statistically significant, LR $\chi^2(26) = 1826.82$, $p < .001$, pseudo-$R^2 = .151$. As shown in Figure 1, the findings suggest that the number of generated requirements varies depending on the LLM used, the contextual knowledge provided, and interaction effects with expert roles. Overall, the amount of requirements generated differ mostly on the contextual knowledge and model (i.e., LLM) utilized. For instance, the no-knowledge condition resulted in significantly more requirements in comparison to the literature condition (IRR ≈ 3.36, $p < .001$). Within the no-knowledge condition, Gemini-3-Pro generated fewer requirements than Claude-Sonnet-4.5 (IRR ≈ 0.25, $p = .027$), whereas GPT-5.2-instant did not significantly differ from Claude-Sonnet-4.5 (IRR ≈ 0.76, $p = .136$). Descriptive statistics show that in the no-knowledge condition Claude-sonnet-4.5 produced the largest and most variable requirement sets ($M = 92.60$, $SD = 48.32$), followed by GPT-5.2-instant ($M = 66.12$, $SD = 22.18$), while Gemini-3-Pro generated substantially smaller sets ($M = 22.31$, $SD = 5.86$). Additional outputs and detailed results are available in the OSF (https://shorturl.at/VLC4z).

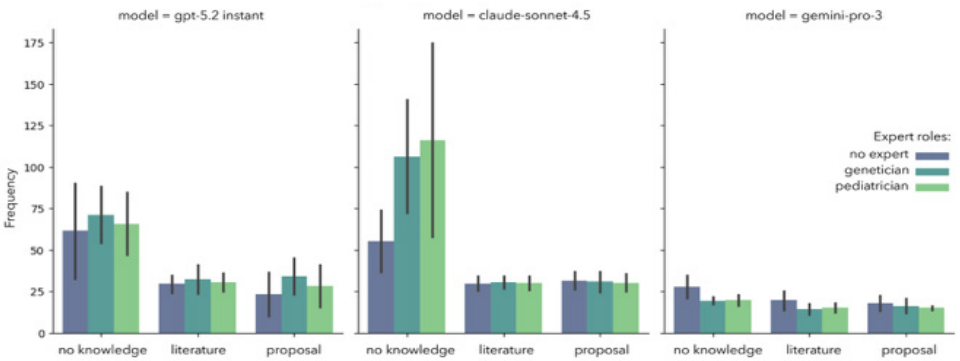

Figure 1: Frequency of requirements per condition.

Average Percentage of Generated Requirements Similar With at Least One Human Expert Requirements

Across conditions, overlap with the human expert generated requirements remained low. For GPT-5.2-instant, overlap ranged from 0.58% to 2.32%, with the highest value observed in the literature condition without an expert role ($M = 0.023$, $SD = 0.026$). Similarly, Claude-sonnet-4.5 range within 0.83% to 2.74%, with the best condition being the literature with the pediatrician role ($M = 0.027$, $SD=0.030$). In contrast, Gemini produced the highest overlap across conditions, ranging from 1.37% to 5.33%. Its best performance occurred under the geneticist role without added knowledge ($M = 0.053$, $SD = 0.042$), followed by the literature knowledge condition ($M = 0.050$, $SD = 0.043$).

Similarity Within-Conditions

The results assessing within-condition reliability indicate that the generated outputs were relatively stable across conditions. All models produced similar requirements across knowledge and persona conditions (mean cosine similarity ≈ 0.95–1.00). A relevant exception was Claude in the literature-only condition, where variability increased under the geneticist role (M = .960, SD = .121, n = 50). In comparison, the most stable results were observed for Claude Sonnet 4.5 when proposal-based knowledge was provided without assigning an explicit expert role (M = .989, SD = .004, n = 50). Overall, the findings suggest consistent reliability across conditions, with mean similarities close to .98 and minimal dispersion.

DISCUSSION

We explored whether providing contextual augmentation and role prompting the gulf of execution and the gulf of evaluation in LLM-supported requirements generation. The key finding suggests that while some conditions increased output volume, they did not improve semantic alignment with human generated requirements. These findings suggests that an execution gap remains, LLMs can generate plausible requirements but do not operationalize contextual knowledge in ways that reflect expert reasoning. The no-knowledge condition generated 3.5 times more requirements than the literature condition, with Claude producing the largest and most variable sets (SD = 48.32). Yet higher quantity did not yield better alignment. Semantic similarity was high across conditions, suggesting that output volume and quality are largely independent. From Norman's perspective, LLMs struggle to translate context into appropriate requirements (Subramonyam et al., 2024). We would expect a closer approximation between LLM-generated requirements and expert-derived requirements. This pattern has practical implications. If the goal is broad coverage i.e., forming an extensive net to capture potential requirements that might otherwise be overlooked, then LLMs without any context may be fine. However, the high variability in Claude's no-knowledge outputs (SD = 48.32) indicates that such breadth comes at the cost of predictability. Requirements engineers using this approach would need thoughtful filtering to manage the noise.

Surprisingly, the inclusion of retrieval-augmented generation did not improve alignment with human-expert requirements. While models supplied with proposal or literature-based knowledge generated fewer requirements, this reduction did not translate into higher similarity scores to the human generated requirements. This result may partly reflect the limitations of similarity metrics, which capture semantic overlap but overlook the pragmatic and structural dimensions that distinguish expert requirements from plausible alternatives. Additionally, the human-expert requirements used as a benchmark were developed through an 18-month iterative process involving workshops, surveys, and negotiation. Such processes integrate tacit knowledge, stakeholder priorities, and contextual judgment that extend beyond what is represented in documents. Even when contextual knowledge is integrated into LLMs, these tools can remain disconnected from the human expert interpretive layer. Consistent with prior work (Quattrocchi

et al., 2025), contextual knowledge in LLMs may constrain outputs without improving their substantive quality. Consequently, RAG may currently serve primarily as a focusing mechanism rather than a quality-enhancing approach in requirements elicitation.

Across all conditions, models demonstrated high within-condition reliability, suggesting that LLM-generated requirements were stable and reproducible across repeated runs. This level of consistency suggests that the models repeatedly produce similar outputs when exposed to the same prompting conditions. However, such stability also points to a tendency for models to converge on a relatively narrow portion of the requirement space rather than exploring a wide range of alternative formulations. In this context, high reliability appears to reflect a focus on a limited set of recurring aspects, rather than diversity in the generated requirements. A similar pattern emerges when considering the effect of role prompting. For the specific task of requirements elicitation, assigning expert roles (e.g., paediatrician or geneticist) had only a marginal effect on the results. Although certain role–knowledge combinations achieved the highest similarity scores, for example GPT-5.2-instant under the paediatrician role, no significant effect was observed for role-prompted conditions overall. One explanation may lie in the limitations of persona-based prompting. While LLMs can emulate role language characteristics associated with a given role, they cannot reproduce the deeper decision-making heuristics, professional priorities, or tacit knowledge that shape how domain experts formulate requirements in practice (Fuentes-Fernández et al., 2010; Hu & Collier, 2024; Uchitel et al., 2022).

These findings suggest a potential supporting role for large language models (LLMs) in early-stage requirements engineering. Even without contextual knowledge augmentation, LLMs can generate a diverse set of requirements at marginal cost. Although these synthetic requirements cannot replace expert judgment, they may serve as conversational scaffolds that help prime discussions, surface overlooked considerations, and reduce the cognitive burden on stakeholders during workshops and focus groups. In this way, LLMs may accelerate early divergent-thinking phases of requirements elicitation. Future work should explore whether richer information injection or more elaborate persona constructions strengthen these effects.

Additionally, future work should take into account some methodological considerations. For instance, in our experiments, we used the POE AI platform, which aggregates access to multiple provider APIs. This might apply to platform-level system prompts or filters that differ across providers, introducing an uncontrolled variable in cross-model comparisons. For this reason, future work could replicate these findings using direct API access under identical configurations. Moreover, the finding that contextual augmentation did not improve alignment could be reflecting the way the contextual knowledge was embedded. In the future, other strategies, such as chunking the information or adding the relevant information to be retrieved in the prompt, can be tested.

CONCLUSION

Large language models can contribute to early-stage requirements elicitation, but not as initially expected. Contextual knowledge augmentation can actually constrains output without improving alignment to human-expert generated requirements. Similarly, prompting LLMs with an expert-role in mind prompting does not imply higher quality requirements. The value of LLM-generated requirements may lie in their capacity to support and decrease the effort of setting up participatory processes. Importantly, using LLMs in requirements engineering requires recognizing that they support generation, not expert decision-making.

FUNDING AND ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101137423. We would like to thank the members of the PROTECT-CHILD consortium for their contributions, insights, and collaboration throughout this project.

REFERENCES

Abo El-Enen, M., Saad, S., & Nazmy, T. (2025). A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Computing and Applications*, *37*(33), 28191–28267. https://doi.org/10.1007/s00521-025-11666-9

Alhoshan, W., Ferrari, A., & Zhao, L. (2025). *How effective are generative large language models in performing requirements classification?* (arXiv:2504.16768). arXiv preprint. https://doi.org/10.48550/arXiv.2504.16768

Anthropic. (2025). *Introducing Claude Sonnet 4.5*. Anthropic News. https://www.anthropic.com/news/claude-sonnet-4-5

Arora, C., Grundy, J., & Abdelrazek, M. (2023). *Advancing requirements engineering through generative AI: Assessing the role of LLMs* (arXiv:2310.13976). arXiv. https://doi.org/10.48550/arXiv.2310.13976

Carley, K., & Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces*, *70*(3), 601–636. https://doi.org/10.2307/2579746

Castle, A. M. R., Goldsmith, C., & Lazier, J. (2025). Working together: Development of a genetic counselling curriculum in a medical genetics residency training program. *Journal of Community Genetics*, *16*(3), 283–289. https://doi.org/10.1007/s12687-025-00798-z

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). *SemEval-2017 Task 1: Semantic textual Similarity - Multilingual and cross-lingual focused evaluation* (arXiv:1708.00055v1). arXiv preprint. https://doi.org/10.18653/v1/S17-2001

Chartier, M., Dakkoune, N., Bourgeois, G., & Jean, S. (2025). HiBenchLLM: Historical Inquiry Benchmarking for Large Language Models. *Data & Knowledge Engineering*, *156*, 102383. https://doi.org/10.1016/j.datak.2024.102383

Chen, G., Yu, Z., Xie, Y., Liu, Z., & Yu, C. (2025). The study of human-AI Co-creation design under generative artificial intelligence: Cognition, process, method, and outcome. *Journal of Engineering Design*, *0*(0), 1–42. https://doi.org/10.1080/09544828.2025.2567155

Cutting, E., Banchero, M., Beitelshees, A. L., Cimino, J. J., Fiol, G. D., Gurses, A. P., Hoffman, M. A., Jeng, L. J. B., Kawamoto, K., Kelemen, M., Pincus, H. A., Shuldiner, A. R., Williams, M. S., Pollin, T. I., & Overby, C. L. (2016). User-centered design of multi-gene sequencing panel reports for clinicians. *Journal of Biomedical Informatics*, *63*, 1–10. https://doi.org/10.1016/j.jbi.2016.07.014

Cysneiros, L. M. (2002). Requirements engineering in the health care domain. *Proceedings IEEE Joint International Conference on Requirements Engineering*, 350–356. https://doi.org/10.1109/ICRE.2002.1048548

Danahey, K., Borden, B. A., Furner, B., Yukman, P., Hussain, S., Saner, D., Volchenboum, S. L., Ratain, M. J., & O'Donnell, P. H. (2017). Simplifying the use of pharmacogenomics in clinical practice: Building the genomic prescribing system. *Journal of Biomedical Informatics*, *75*, 110–121. https://doi.org/10.1016/j.jbi.2017.09.012

Fuentes-Fernández, R., Gómez-Sanz, J. J., & Pavón, J. (2010). Understanding the human context in requirements elicitation. *Requirements Engineering*, *15*(3), 267–283. https://doi.org/10.1007/s00766-009-0087-7

Google. (n.d.). *Gemini 3 Pro | Generative AI on Vertex AI | Google Cloud Documentation*. Google Cloud Documentation. Retrieved January 19, 2026, from https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-pro

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

Hu, T., & Collier, N. (2024). *Quantifying the persona effect in LLM simulations* (arXiv:2402.10811). arXiv preprint. https://doi.org/10.48550/arXiv.2402.10811

Karolita, D., Grundy, J., Kanij, T., Obie, H., & McIntosh, J. (2024). CRAFTER: A persona generation tool for requirements engineering: *Proceedings of the 19th International Conference on Evaluation of Novel Approaches to Software Engineering*, 674–683. https://doi.org/10.5220/0012718400003687

Linåker, J., Regnell, B., & Damian, D. (2020). A method for analyzing stakeholders' influence on an open source software ecosystem's requirements engineering process. *Requirements Engineering*, *25*(1), 115–130. https://doi.org/10.1007/s00766-019-00310-3

Meyer, J., Ostrzinski, S., Fredrich, D., Havemann, C., Krafczyk, J., & Hoffmann, W. (2012). Efficient data management in a large-scale epidemiology research project. *Computer Methods and Programs in Biomedicine*, *107*(3), 425–435. https://doi.org/10.1016/j.cmpb.2010.12.016

Mouri Zadeh Khaki, A., Choi, A., & Seyyed-Kalantari, L. (2025). Simulating social behavior of LLM-based autonomous negotiator agents in a game-theoretical framework using multi-agent systems. *International Journal of Human–Computer Interaction*, *41*(23), 15169–15178. https://doi.org/10.1080/10447318.2025.2495117

Norman, D. A. (1992). Design principles for cognitive artifacts. *Research in Engineering Design*, *4*(1), 43–50. https://doi.org/10.1007/BF02032391

OpenAI. (n.d.). *Using GPT-5.2 | OpenAI API*. OpenAI API Documentation. Retrieved January 19, 2026, from https://platform.openai.com/docs/guides/latest-model

Quattrocchi, G., Pasquale, L., Spoletini, P., & Baresi, L. (2025). *Can LLMs generate user stories and assess their quality?* (arXiv:2507.15157). arXiv preprint. https://doi.org/10.48550/arXiv.2507.15157

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence embeddings using siamese BERT-networks* (arXiv:1908.10084). arXiv preprint. https://doi.org/10.48550/arXiv.1908.10084

Sharp, H., Finkelstein, A., & Galal, G. (1999). Stakeholder identification in the requirements engineering process. *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99*, 387–391. https://doi.org/10.1109/DEXA.1999.795198

Sobo, A., Mubarak, A., Baimagambetov, A., & Polatidis, N. (2025). Evaluating LLMs for Code Generation in HRI: A Comparative Study of ChatGPT, Gemini, and Claude. *Applied Artificial Intelligence*, *39*(1), 2439610. https://doi.org/10.1080/08839514.2024.2439610

Sommerville, I. (2016). *Software engineering* (10th ed.). Pearson Education, Inc.

Subramonyam, H., Pea, R., Pondoc, C., Agrawala, M., & Seifert, C. (2024). Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3613904.3642754

Tehenan, M. (2025). Semantic geometry of sentence embeddings. *Findings of the Association for Computational Linguistics: EMNLP 2025.*, 11993–12004.

Uchitel, J., Alden, E., Bhutta, Z. A., Cavallera, V., Lucas, J., Oberklaid, F., Patterson, J., Raghavan, C., Richter, L., Rikard, B., Russell, R. R., & Mikati, M. A. (2022). Role of pediatricians, pediatric associations, and academic departments in ensuring optimal early childhood development globally: Position paper of the International Pediatric Association. *Journal of Developmental & Behavioral Pediatrics*, *43*(8), e546–e558. https://doi.org/10.1097/DBP.0000000000001112

Van Velsen, L., Wentzel, J., & Van Gemert-Pijnen, J. E. (2013). Designing eHealth that matters via a multidisciplinary requirements development approach. *JMIR Research Protocols*, *2*(1), e21. https://doi.org/10.2196/resprot.2547