A Reference Model for Empirically Comparing LLMs with Humans

Kurt Schneider
Software Engineering
Leibniz Universität Hannover
Hannover, Germany
kurt.schneider@inf.uni-hannover.de

Farnaz Fotrousi
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden
farnaz.fotrousi@gu.se

Rebekka Wohlrab
Chalmers University of Technology
University of Gothenburg
Carnegie Mellon University
Gothenburg, Sweden and Pittsburgh, PA, USA
wohlrab@chalmers.se

***Abstract*—Large Language Models (LLM) have shown stunning abilities to carry out tasks that were previously conducted by humans. The future role of humans and the responsibilities assigned to non-human LLMs affect society fundamentally. In that context, LLMs have often been compared to humans. However, it is surprisingly difficult to make a fair empirical comparison between humans and LLMs. To address those difficulties, we aim at establishing a systematic approach to guide researchers in comparing LLMs with humans across various tasks. In a literature review, we examined key differences and similarities among several existing studies. We developed a reference model of the information flow based on that literature exploration. We propose a framework to support researchers in designing and executing studies, and in assessing LLMs with respect to humans. Future studies can use the reference model as guidance for designing and reporting their own unique study design. We want to support researchers and the society to take a maturation step in this emerging and constantly growing field.**

***Index Terms*—LLM, Reference Model, empirical evaluation**

I. Introduction

LLMs understand and create natural language, and their reactions resemble potential reactions of humans. In many cases, the output of an LLM looks as if it came from a human. There are good reasons to believe that an LLM might react like a human: LLMs are trained on human-created sources and they tend to replicate human behavior in certain ways.

LLMs *are* indeed models of humans. In his model theory, Stachoviak [22] asks for (1) the *original*, (2) the *receiver* and *purpose* of a model, and (3) the resulting *relevant aspects* for that purpose. In our case, "humans" can be seen as the original. LLMs are typically compared to humans for the following reasons and purposes:

- Feasibility: Demonstrate that a difficult task can be performed without humans involved.
- Quality criterion: Use humans as a gold standard for the achievements of an LLM.
- Replace specific humans by an LLM.

LLMs are leveraged for a variety of reasons and in a variety of contexts. A company may consider hiring humans—or replacing them with an LLM. Non-native students in school may benefit from LLMs supporting them when writing English texts. Telephone health advisors can act empathically and interact with humans who describe their problems. Humans can be replaced in many repetitive tasks. Society, on the other hand, might be interested in detecting bias in a decision. All aspects of LLM-as-a-model (original, receiver, purpose, relevant aspects) affect the evaluation of a given LLM [2]. Guo et al. present a comprehensive survey on evaluating LLMs [8].

Comparing LLMs to humans raises questions about the consequences and ethical implications for society: Is it ethically acceptable for an LLM to pretend being a human—assuming that this mimikry helps users to accept its recommendations? On the other hand, LLM may detect deeply human misbehavior more objectively than humans (e.g., identifying hate speech), as Huang et al. found [11]. Societal stakeholders and researchers need to understand what a comparison of LLM with humans really means in each specific case – and how to interpret and evaluate the results; e.g., are humans (or entire populations [17]) to be supported or rather simulated [7]?

Empirically comparing LLMs with humans is difficult - which creates a problem for both researchers and readers.

Why a comparison is difficult: Challenges start with providing "the same task" to LLMs and to humans. A description for a human domain expert can assume some familiarity with specific needs and requirements in the domain. An LLM has all the information on the internet, but must be informed via a prompt about the specific situation. Salewski recommends informing an LLM of its own role [24], which a human will already know. Simply using the same task description for a human and an LLM will often be unfair to one or the other.

Diversity: There is not "one typical human", nor is there just one LLM that answers in a consistent way, as Atari et al. highlight in their paper entitled: "Which Humans?" [3]. The probabilistic nature of LLMs and the diversity of humans call for statistical means in comparing "LLMs" with "humans".

What is better? What does it mean in statistical terms for an LLM to be "better than humans" [11]: Is the LLM always better than any human? Or better on average than average human results? Or is a given LLM mostly better or never worse than humans, according to a given metric?

II. Research Method

Fig. 1 shows the steps of our research method. Our study was motivated by existing publications that compare LLMs with humans for specific tasks, as described in Sect. I. We encountered similar challenges in our own empirical studies.

① Initially, we conducted a literature review, focusing on papers from the past 4 years. We intended to start by exploring the topic instead of providing a full mapping study. Therefore, we restricted our findings to the first 100 hits in Google Scholar, using the search string *"(LLM OR ChatGPT) AND (human or person) AND (empirical or compar)"*, where *compar* matched with compare as well as with comparison. We systematically filtered the results by applying *inclusion and exclusion criteria* to title, abstract, and –when in doubt – to the full papers. This led to 12 (out of 100) papers being included: [2], [3], [5], [8], [12], [18], [24], [26], [28], [30], [32], [33]; 7 are arXiv papers. We included arXiv papers to see how researchers report recent their studies about this very new subject. After that, we conducted one round of forward snowballing, i.e., we searched for all papers that had cited one of the 12 initial articles, which led to another 14 articles (incl. 11 arXiv): [1], [4], [6], [7], [9], [10], [13], [15]–[17], [20], [21], [25], [31].

② We then read the 12+14 papers informally. We noticed how heterogeneously the comparisons were presented. At that point, we decided to collect, compare, and visualize crucial activities and decisions in empirical studies of humans vs. LLMs. Finally, we decided to provide a reference model to put all these aspects in context.

③ Deriving the reference model. *Step 3.1:* We analyzed one paper after the other, collecting and classifying stated goals and respective comparisons. Goals consisted of a claim (classified as: Feasibility, Quality, Replacement; see Introduction) and an explanation of the task at hand (e.g., create Java code). *Step 3.2:* Next, we checked how each aspect of achieving a goal was measured and, *Step 3.3:*, how measurements were compared between humans and LLMs. There were different ways of comparing the measurements, ranging from simple numerical or statistical comparisons to expert evaluations, benchmarks, and sophisticated schemes developed for a given task. We also noticed the different ways of communicating "the task" to LLM and human, respectively. *Step 3.4, Iterative Abstraction:* We abstracted from specific examples in a study. When a new paper was analyzed, we tried to map its parts to the previously identified aspects and decisions from steps 3.1-3.3. In some cases, this mapping did not work, due to a missing or inadequate aspect. In these cases, we refined or adapted the model. For example, we split "the comparison activity" into a specific measurement for the human and for the LLM sides and then a comparison between the two measurements when it became clear that LLMs and humans were often measured differently. This also emphasized the need to specify exactly how the comparison was conducted - and how this specific comparison relates to the initially stated goal of a publication.

From a methodological point of view, *our process of deriving the reference model was an iterative comparison and abstraction of elements in the papers*. The abstraction serves a similar purpose as coding in other contexts but our abstraction is not as fine-grained as e.g., open coding. It also considers the fact that the presentations in different papers vary significantly.

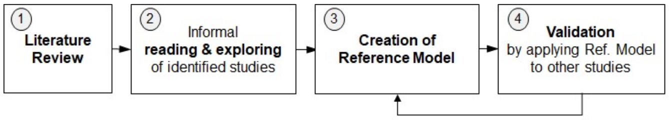
①
Literature Review
②
Informal reading & exploring of identified studies
③
Creation of Reference Model
④
Validation by applying Ref. Model to other studies

Fig. 1. Overview of our research method

Step 3.5, Depencencies and Information Flow: Obviously, one aspect (e.g., description of task) can influence others (e.g., what aspect of the result is relevant). In addition, many aspects need to be reported explicitly rather than implicitly (e.g., what exactly constitutes the Ground Truth). We decided to visualize dependencies in a graphical model. This model should be able to show the flow of information and decisions in a study from one activity to the next. It should show all relevant documents (explicit information), but also information that is often being communicated verbally – and, hence, does not appear explicitly in a paper. Key decision points identified in previous steps should be visible in this overview diagram.

Step 3.6, Choice of notation: The FLOW notation [27] meets these requirements. Fig. 2 shows the current version of our reference model as FLOW model.

④ To indicate how the reference model can be validated, we applied it to a different set of papers. They should represent diverse publications about LLMs vs. humans to see if it is feasible to map other studies to the reference model. We demonstrate this approach by selecting three related papers randomly. A larger sample will need to be checked systematically – which is beyond the scope and space of this paper. We expect to find additional relevant aspects of studies when we extend the literature search (arrow pointing from ④ to ③).

III. Reference Model

A. Purpose of a Reference Model

We created a reference model to put our findings in context. A reference model can serve various purposes:

- As a checklist for researchers *planning* a study: what decisions must be taken, how does information flow between the steps of the study?
- As a checklist for *reporting* on empirical studies: avoiding unclear and implicit important assumptions and decisions.
- As an aid to *classify and categorize* studies: What are commonalities and differences between given studies in terms of key decisions?

B. The FLOW Notation

Fig. 2 shows our reference model in FLOW notation. The notation is intentionally kept simple. It is defined in [27]. This notation visualizes both document-based and informal, verbal flows of information. *Document symbols* represent explicit information, solid arrows indicate its flow through the study. Word or PDF documents are examples, as well as videos. *Implicit information and verbal flow* typically originate from humans: They talk or scribble to convey information. A face symbol (smiley) represents a storage, a dashed arrow indicates

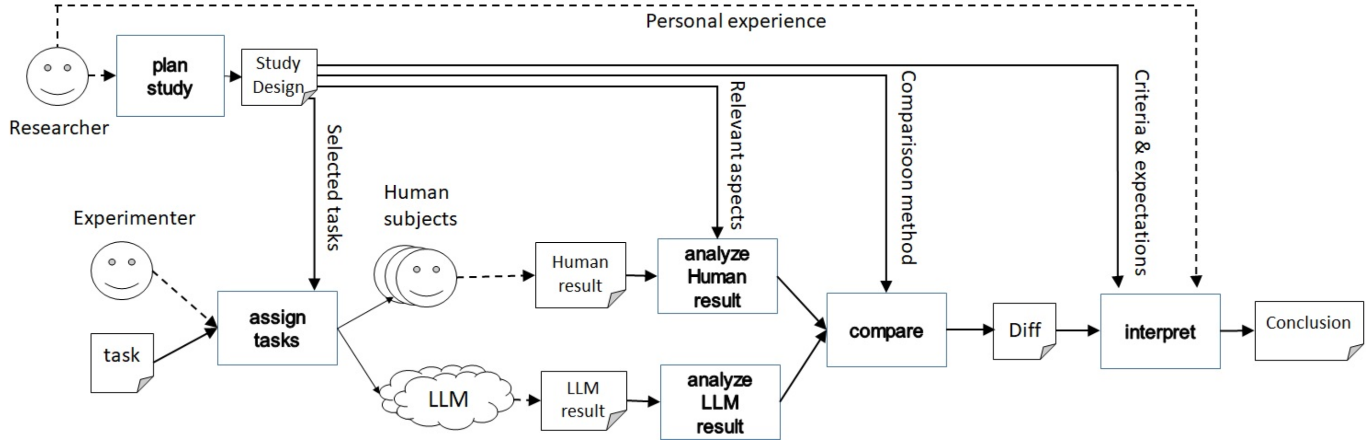

Fig. 2. Reference Model for empirical studies comparing LLMs with humans. Information flows along arrows and activities to finally reach the Conclusion. A new study can use this model to map its own parts to the documents, people and activities shown in this reference (see also explanation in III.C).

a flow of this type. Verbally communicated information that is not made explicit (and documented) flows faster but will soon be forgotten and can rarely be recovered precisely; in particular, implicit decisions or assumptions in a paper are difficult to reconstruct. *Boxes indicate activities*, with incoming and outgoing flows. Control flows (e.g., by a checklist) come in from the top. *LLMs* are neither human nor documents. They are, therefore, depicted as cloud symbols with dashed arrows, since LLMs often do not create the same output when asked again later (similar to verbally communicated information).

C. Key Decisions and Information Flow in Empirical Studies

The reference model in Fig. 2 is a compilation of decisions and information flows extracted from existing empirical papers, as described in Section II. They are now ordered and placed in Fig. 2 to show dependencies and information flow:

1) What is the *goal or purpose* of the study? How will results be used and what are potential consequences of findings? These crucial decisions are needed to plan the study. The following decisions build on them.
2) *Dependencies and flow:* Decisions and information flow must be planned from top to bottom (goal of the study to instantiation) and from the end (desired outcome) to the beginning (required input). Results and findings are then collected, interpreted, and evaluated during the study.
3) *Criteria and expectations:* In the end, the actual findings of a study are put in perspective of its initial goals and expectations. Although this information is used only at the end of the study to evaluate findings, it must be decided and reported early in the Study Design activity.
4) *Comparison method*: How can human results and LLM results be effectively compared to reach the goals of a study? In terms of Stachoviak's model theory: what are *relevant aspects* of the results? What are adequate quantitative or qualitative techniques to compare them?
5) *Analyze for comparison*: In many studies, a result is pre-analyzed (e.g., counted, coded, computed) before it is compared to the other result (i.e., stemming from LLM vs. human). A *comparison* is typically conducted on those (pre-analyzed) results of a task.
6) As discussed above, it is difficult to assign and phrase *task descriptions* fairly. Selecting a task and assigning it (in "assign tasks") constitutes a crucial decision that must take all earlier decisions into account.
7) In the final *interpretation*, explicit and documented criteria are applied, compared to pre-defined expectations, and finally reflected and described in the light of the personal experience of a researcher and author.

D. Applying the Reference Model

Different studies vary substantially in reporting the above-mentioned items, which is good: Authors and researchers must remain creative in planning their studies. However, it is to the advantage of both researchers and readers to report clearly what and how they found it, and what can be concluded. The reference model can be an inspiration and checklist:

- *Planning*: Authors can **check** with respect to Fig. 2 if they already carried out all activities shown in the reference model and answered all questions in Sec. III.C.
- *Reporting*: Are decisions in "plan study" **documented** for reproducible interpretation (rather than implicit)?
- *Refining*: How are all activities conducted in detail? Are the details of collecting data compliant with the chosen **Comparison method**, and will this method create a description of the **Diff**(erences) that enables the desired type of **Conclusions**?
- *Comparing* studies: If two or more studies map their decisions and information flow to the reference model, their profiles (activities, flows, decisions) can be compared. E.g., a selected task might be the same but comparison techniques may differ due to different purposes.

IV. Validation Approach

This section presents the suggested validation approach for the reference model, based on an in-depth analysis of a first, small sample of three randomly selected software engineering

TABLE I
MAPPING THREE STUDIES TO THE REFERENCE MODEL.

Paper	***Purpose***	***LLM vs Human***	***Selected Tasks***	***Analysis of results***	***Comparison method***	***Interpret***
[29]	Purpose-A*	ChatGPT 3.5	Coding puzzles or development tasks with ChatGPT's assistance	Measure efficiency, solution quality, subjective perception and task load	Mann Whitney U test, calculate effect size	Interpret-A*
		Human	Coding puzzles or development tasks without ChatGPT's assistance			
[14]	Purpose-B*	ChatGPT 3.5 and 4.0	102 challenges from the IEEExtreme	Calculate a Score with the inputs such as incomplete code produced, and compile errors	Compare ChatGPT's score with average score of humans	Interpret-B*
		Human	Benchmark of the 102 challenges			
[19]	Purpose-C*	Diverse LLMs	Using the engineered prompts, label data for three NLP tasks	Measure Precision, recall, F1, Inter-Annotator Agreement	ANOVA for analysis of variance and Cohen's d for Effect Size	Interpret-C*
		Human	Using a comprehensive guideline, label data in three NLP tasks			

*: Stated within the text

papers [29] [14] [23]. These papers were chosen to represent diverse profiles of publications comparing humans and LLMs. For this initial validation, we extracted key elements used in designing and executing each study, and mapped them to the reference model. The validation should check if this can be done effectively. In the following, we present the mappings of the studies to the reference model in Tab. I.

Purpose-A of study A [29] was to investigate how helpful it was to work with ChatGPT in software development tasks. In a controlled experiment, researchers *assigned* coding puzzles and typical software development **tasks** to four groups of participants to perform their tasks either by themselves as *humans* or getting support from *ChatGPT*. Efficiency (how fast the task was solved) was among the *relevant aspects*, besides solution quality, subjective perception, and task load. The **analysis** showed that those working with ChatGPT took 2196.7 sec. vs. 2692.6 sec. for those without it. As *comparison method*, the differences between those durations were statistically **compared** by the M-U Wilcoxon test and found significant, (p-value = 0.026, effect r = 0.30). The **interpretation** (i.e., Interpret-A) saw significant improvements in efficiencies for using ChatGPT in coding puzzles but only slight and insignificant improvements for development tasks.

Purpose-B of study B [14] was to investigate ChatGPT's problem-solving capabilities in programming tasks. The study involved solving the **task** of 102 challenges from the programming competition IEEExtreme. The challenges were assigned to *ChatGPT* via prompts and compared against a *benchmark* of performance created by *humans*. The **analysis** found the average performance score of ChatGPT on Python tasks to be 9.06 vs. 44.5 for humans. This score contains criteria like incomplete code, errors, and failed test cases. The numerical **comparison** showed that the average human score was 3.9 to 5.8 times higher than of ChatGPT, **interpreted** (i.e., Interpret-B) as humans' superior problem-solving skills in programming.

Purpose-C of study C [19] was to investigate whether LLMs can effectively compete with humans in annotating data for sentiment analysis. The study examined NLP annotation **tasks**—topic classification, sentiment analysis, and emotion classification—in three languages, using guidelines for humans and task-specific prompts for LLMs. In the **analysis**, annotation accuracy, precision, recall, and F1-score were measured. The **comparison** of results showed a Cohen's d of 0.6 in the emotion classification task. **Interpretation** (i.e., Interpret-C) indicates human annotators consistently outperformed in more complex tasks such as emotion classification while LLMs performed competitively in simpler tasks.

Extracting relevant information from existing papers was difficult, as the information was often unstructured or implicit. The above three examples demonstrate that different software engineering studies can be mapped to the reference model, and they illustrate what the mappings can look like. We suggest continuing validation following this critical mapping approach.

V. DISCUSSION

This paper proposes a reference model for empirical studies that compare the performance of LLMs with humans. The model was derived in an iterative refinement process by abstracting key elements from 12+14 papers identified in keyword search and snowballing. The model is visualized to show dependencies and the bigger picture. Model elements and key questions are described in III.C.

We suggest mapping other papers to that reference model, to allow planning, reporting, and comparison of new studies. We demonstrate how the reference model can be validated and evolved. Diversity of humans must be considered [28] as well as clear criteria for analysis, comparison, and interpretation.

We envision the reference model to be also used for assessing human-LLM collaboration and teaming. Informed interpretation and conclusions of comparing humans, LLMs, and their collaboration will be crucial for the development of software engineering and for the future of society.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Abbasiantaeb, Z., Yuan, Y., Kanoulas, E., Aliannejadi, M.: Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. In: Proc. of the Int. Conf. on Web Search and Data Mining. pp. 8–17 (2024)

[2] Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: Proc. of the Int. Conf. on Machine Learning. pp. 337–371. PMLR (2023)

[3] Atari, M., Xue, M.J., Park, P.S., Blasi, D., Henrich, J.: Which humans? PsyArXiv Preprints (2023)

[4] Awasthi, R., Mishra, S., Mahapatra, D., Khanna, A., Maheshwari, K., Cywinski, J., Papay, F., Mathur, P.: HumanELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. medRxiv pp. 2023–12 (2023)

[5] Chen, C., Yao, B., Ye, Y., Wang, D., Li, T.J.J.: Evaluating the LLM agents for simulating humanoid behavior. In: Proc. of the 1st Workshop on Human-Centered Evaluation and Auditing of Language Models (CHI Workshop HEAL) (2024)

[6] Dominguez-Olmedo, R., Hardt, M., Mendler-Dünner, C.: Questioning the survey responses of large language models. arXiv preprint arXiv:2306.07951 (2023)

[7] Gui, G., Toubia, O.: The challenge of using LLMs to simulate human behavior: A causal inference perspective. arXiv preprint arXiv:2312.15524 (2023)

[8] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., Wu, Y.: How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597 (2023)

[9] Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Yu, L., Liu, Y., Li, J., Xiong, B., Xiong, D., et al.: Evaluating large language models: A comprehensive survey. arXiv preprint arXiv:2310.19736 (2023)

[10] Hu, T., Collier, N.: Quantifying the persona effect in LLM simulations. arXiv preprint arXiv:2402.10811 (2024)

[11] Huang, F., Kwak, H., An, J.: Is ChatGPT better than human annotators? potential and limitations of ChatGPT in explaining implicit hate speech. In: Proc. of the ACM Web Conf. Companion 2023. pp. 294–297 (2023)

[12] Jiang, G., Xu, M., Zhu, S.C., Han, W., Zhang, C., Zhu, Y.: Evaluating and inducing personality in pre-trained language models. Advances in Neural Information Processing Systems **36** (2024)

[13] Kabir, S., Udo-Imeh, D.N., Kou, B., Zhang, T.: Is stack overflow obsolete? an empirical study of the characteristics of ChatGPT answers to stack overflow questions. In: Proc. of the 2024 CHI Conference on Human Factors in Computing Systems. pp. 1–17 (2024)

[14] Koubaa, A., Qureshi, B., Ammar, A., Khan, Z., Boulila, W., Ghouti, L.: Humans are still better than ChatGPT: Case of the IEEEXtreme competition. Heliyon **9**(11) (2023)

[15] Lin, Z.: Large language models as probes into latent psychology. CoRR (2024)

[16] Ma, Q., Xue, X., Zhou, D., Yu, X., Liu, D., Zhang, X., Zhao, Z., Shen, Y., Ji, P., Li, J., et al.: Computational experiments meet large language model based agents: A survey and perspective. arXiv preprint arXiv:2402.00262 (2024)

[17] Namikoshi, K., Filipowicz, A., Shamma, D.A., Iliev, R., Hogan, C.L., Arechiga, N.: Using LLMs to model the beliefs and preferences of targeted populations. arXiv preprint arXiv:2403.20252 (2024)

[18] Nascimento, N., Alencar, P., Cowan, D.: Comparing software developers with ChatGPT: An empirical investigation. arXiv preprint arXiv:2305.11837 (2023)

[19] Nasution, A.H., Onan, A.: ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. IEEE Access (2024)

[20] Ng, M., Tse, H.T., Huang, J., Li, J., Wang, W., Lyu, M.R.: How well can LLMs echo us? evaluating AI chatbots' role-play ability with ECHO. arXiv preprint arXiv:2404.13957 (2024)

[21] Nguyen, T.H., Rudra, K.: Human vs ChatGPT: Effect of data annotation in interpretable crisis-related microblog classification. In: Proc. of the 2024 ACM Web Conf. pp. 4534–4543 (2024)

[22] Niemeyer, K.: A contribution to model theory. NATO Security through Science Series - D: Information and Communication Security **12**, 25 (2007)

[23] Rodriguez-Echeverría, R., Gutiérrez, J.D., Conejero, J.M., Prieto, Á.E.: Analysis of ChatGPT performance in computer engineering exams. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje (2024)

[24] Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., Akata, Z.: In-context impersonation reveals large language models' strengths and biases. Advances in Neural Information Processing Systems **36** (2024)

[25] Shi, Z., Wang, Z., Fan, H., Zhang, Z., Li, L., Zhang, Y., Yin, Z., Sheng, L., Qiao, Y., Shao, J.: Assessment of multimodal large language models in alignment with human values. arXiv preprint arXiv:2403.17830 (2024)

[26] Sicilia, A., Gates, J.C., Alikhani, M.: HumBEL: A human-in-the-loop approach for evaluating demographic factors of language models in human-machine conversations. arXiv preprint arXiv:2305.14195 (2023)

[27] Stapel, K., Schneider, K.: Managing knowledge on communication and information flow in global software projects. Expert Systems **31**(3), 234–252 (2014)

[28] Wang, A., Morgenstern, J., Dickerson, J.P.: Large language models cannot replace human participants because they cannot portray identity groups. arXiv preprint arXiv:2402.01908 (2024)

[29] Wang, W., Ning, H., Zhang, G., Liu, L., Wang, Y.: Rocks coding, not development: A human-centric, experimental evaluation of LLM-supported SE tasks. In: Proc. of the ACM on Software Eng. vol. 1, pp. 699–721. ACM New York, NY, USA (2024)

[30] Welivita, A., Pu, P.: Is ChatGPT more empathetic than humans? arXiv preprint arXiv:2403.05572 (2024)

[31] Yu, Z., Gao, C., Yao, W., Wang, Y., Ye, W., Wang, J., Xie, X., Zhang, Y., Zhang, S.: Kieval: A knowledge-grounded interactive evaluation framework for large language models. arXiv preprint arXiv:2402.15043 (2024)

[32] Zhao, Y., Zhang, J., Chern, I., Gao, S., Liu, P., He, J., et al.: Felm: Benchmarking factuality evaluation of large language models. Advances in Neural Information Processing Systems **36** (2024)

[33] Zhuang, Y., Liu, Q., Ning, Y., Huang, W., Lv, R., Huang, Z., Zhao, G., Zhang, Z., Mao, Q., Wang, S., et al.: Efficiently measuring the cognitive ability of LLMs: An adaptive testing perspective. arXiv preprint arXiv:2306.10512 (2023)