Investigating the Use of LLMs for Evidence Briefings Generation in Software Engineering

* Indicates required question

Consent Form

This form concerns your participation in the study *"Evaluating the Use of LLMs to Automatically Generate Evidence Briefings in Software Engineering."* Your participation is voluntary, and all responses will be anonymous and used solely for research purposes.

To ensure ethical and transparent data collection, this form follows the best practices outlined by Badampudi et al. (2017). You may withdraw at any time without penalty.

Please read the following before continuing.

1. **1. Title and Purpose of the Study** *

You are invited to participate in a research study titled *“Investigating the Use of LLMs for Evidence Briefings Generation in Software Engineering”*.
The goal of this study is to explore how Large Language Models (LLMs) can be used to automatically generate Evidence Briefings—short, practitioner-oriented summaries of research papers. As a researcher, you will be asked to evaluate whether the summaries presented accurately reflect the content of the original studies.

2. What Will Participation Involve?

If you agree to participate, you will complete a short online questionnaire. You will be presented with two Evidence Briefings along with their original research articles. For each, you will assess how accurately the summary reflects the source material. This activity is expected to take approximately 15 to 20 minutes.

3. Voluntary Participation

Your participation is completely voluntary. You may withdraw at any time without any consequences.

4. Risks and Benefits

There are no known or anticipated risks associated with this study. While you will not receive any direct benefits, your participation will contribute to ongoing research efforts aimed at improving knowledge transfer between academia and industry.

5. Anonymity and Confidentiality

We will not collect any personally identifiable information such as your name or email. All data will be analyzed in aggregate form. No individual responses will be reported.

6. Data Storage and Usage

Your responses will be stored securely and used exclusively for academic research purposes. Anonymized data may be included in publications or made available through open-access research repositories. No personally identifiable information will ever be disclosed.

7. Contact Information

If you have any questions about this study, please contact the researchers at [mauro.mmo93@gmail.com].

8. Consent Statement

By selecting "I agree" below, you confirm that:

Are at least 18 years old;

Have read and understood the information presented in this consent form;

Understand that your participation is voluntary and you may withdraw at any time;

Voluntarily agree to participate in this research study.

Mark only one oval.

- ◯ I agree
- ◯ I disagree

Characterization Form

This section collects background information to help us analyze responses according to your profile. It includes questions about your professional experience and familiarity with software engineering research. Your answers will remain anonymous and are essential for ensuring the validity of our analysis.

2. How familiar are you with Systematic Literature Reviews? *

Mark only one oval.

	1	2	3	4	5	
Not	◯	◯	◯	◯	◯	Very familiar

3. What is your current academic role? *(Select the one that best describes your position)* *

Mark only one oval.

- ◯ University Professor(Assistant Professors included)
- ◯ Research Scientist in Industry
- ◯ Post-doc
- ◯ Ph.D. Student
- ◯ MS Student

4. In which country are you based? *

_____

Main Section

In the following section, you will have acess to two research articles and their respective Evidence Briefings. Please read the original research paper in order to answer questions about content fidelity.

The effectiveness of pair-programming: A meta-analysis
https://drive.google.com/file/d/1sJPeSqZ-Mn3bCvXbngsxxu3S-V7oH6v-/view?usp=sharing

THE EFFECTIVENESS OF PAIR PROGRAMMING

This briefing reports evidence on the effectiveness of pair programming around quality, duration, and effort based on scientific evidence from a systematic review.

MAIN FINDINGS

• The findings presented in this briefing consider quality as the number of test cases passed or number of correct solutions of programming tasks; duration as the total time taken to complete all tasks considered (all solutions); and effort was reported as twice the duration of each individual in the pair.

• Studies present a small significant positive overall effect of pair programming on quality, a medium significant positive overall effect on duration, and a medium significant negative overall effect on effort.

• Evidence suggests that pair programming is faster than solo programming when programming task complexity is low and also yields code solutions of higher quality when task complexity is high.

• The higher quality for complex tasks comes at a price of considerably greater effort, while the reduced completion time for the simpler tasks comes at a price of noticeably lower quality.

• Research results show that the question of whether pair programming is better than solo programming depends on other factors, for example, the expertise of the programmers and on the complexity of the system and tasks to be solved.

• One of the most interesting observations is that the pairing up of individuals seems to elevate the junior pairs up to near senior pair performance. Thus, pair collaboration might compensate for juniors' lack of deep understanding, for example, by inducing an expert-like strategy.

• If you do not know the seniority or skill levels of your programmers, but do have a feeling for task complexity, then employ pair programming either when task complexity is low and time is of the essence, or when task complexity is high and correctness is important.

• When considering the moderating effect of programmer expertise, junior pairs had a small (5%) increase in duration and thus a large increase in effort (111%), and a 73% increase in correctness.

• Intermediate pairs had a 28% decrease in duration (43% increase in effort) and a negligible (4%) increase in correctness.

• Senior pairs had a 9% decrease in duration (83% increase in effort) and an 8% decrease in correctness.

• The juniors benefited from pair programming in terms of increased correctness, the intermediates in terms of decreased duration, while there were no overall benefits of pair programming for seniors.

• When considering the combined moderating effect of system complexity and programmer expertise on pair programming, there appears to be an interaction effect: Among the different treatment combinations, junior pairs assigned to the complex design had a remarkable 149% increase in correctness compared with individuals.

• Intermediates and seniors experienced an effect of pair programming on duration on the simpler design, with a 39% and 23% decrease, respectively.

• However, the cost of this shorter duration was a corresponding decrease in correct solutions by 29% and 13%, respectively.

Keywords: Pair programming, Meta-analysis

Who is this briefing for?

Software engineering practitioners who want to make decisions about pair programming based on scientific evidence.

Where the findings come from?

All findings of this briefing were extracted from the systematic review conducted by Hannay et al.

What is included in this briefing?

The main findings of the original systematic review. Evidence characteristics through a brief description about the original systematic review and the studies it analyzed.

What is not included in this briefing?

Additional information not presented in the original systematic review. Detailed descriptions about the studies analyzed in the original systematic review.

For additional information about this briefing: cin.ufpe.br/eseg/briefings

5. To what extent do you agree with the following statement: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Slightly Disagree	I neither agree nor disagree	Slightly Agree	Agree	Strongly Agree
The conclusions presented in the briefing are consistent with those in the original article.	⬭	⬭	⬭	⬭	⬭	⬭	⬭

6. Please justify your choice

7. To what extent do you agree with the following statement: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Slightly Disagree	I neither agree nor disagree	Slightly Agree	Agree	Strongly Agree
The level of certainty expressed in the briefing matches the confidence level presented in the original article.	◯	◯	◯	◯	◯	◯	◯

8. Please justify your choice

9. To what extent do you agree with the following statement: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Slightly Disagree	I neither agree nor disagree	Slightly Agree	Agree	Strongly Agree
All claims made in the briefing are supported by information present in the original article.	◯	◯	◯	◯	◯	◯	◯

10. Please justify your choice

_____

_____

_____

_____

_____

A systematic review on the use of Definition of Done on agile software development projects
https://drive.google.com/file/d/1cCvNpGVsU6aTCl4SSFQKaYcCOvopTeAT/view?usp=sharing

EVIDENCE BRIEFING ON THE USE OF DEFINITION OF DONE IN AGILE SOFTWARE DEVELOPMENT PROJECTS

This briefing reports evidence on the use of the Definition of Done (DoD) in agile software development, based on a systematic review of the existing literature. The goal is to provide practitioners with insights into how DoD is applied in different contexts and to highlight areas for future research.

FINDINGS

- The systematic review analyzed 2,326 papers, ultimately identifying 8 studies that included criteria for DoD in agile projects. This indicates a limited but growing body of literature on this critical aspect of agile practices.

- A total of 62 distinct DoD criteria were identified across the studies, categorized into software verification and validation, deployment processes, code inspection, test quality, regulatory compliance, software architecture, process management, configuration management, and non-functional requirements.

- Some studies utilized a multi-level approach to DoD, encompassing various levels such as story, sprint, release, and project, which allows for tailored criteria depending on the project context.

- The findings suggest that while many teams recognize the importance of DoD, there is significant variability in its application. Only a few criteria, like unit tests and peer code reviews, were consistently reported across multiple studies, indicating a lack of consensus on what constitutes a comprehensive DoD.

- The review highlighted that the quality of the studies was generally low, with many being experience reports rather than rigorous empirical research. This raises concerns about the reliability of the findings and the need for more robust empirical studies to validate the use of DoD in practice.

PRACTICAL IMPLICATIONS

- Practitioners can utilize the findings as a reference to define or refine their own DoD criteria, ensuring that it aligns with their specific project needs and contexts.

- The identified criteria can serve as a checklist to enhance quality management practices within agile teams, potentially leading to improved collaboration and reduced defects.

- Organizations should consider documenting their DoD practices and outcomes, contributing to the growing body of evidence and helping to establish best practices within the agile community.

WHO IS THIS BRIEFING FOR?

This briefing is intended for software engineering practitioners, project managers, and agile coaches who are looking to enhance their understanding and implementation of the Definition of Done in agile software development projects.

WHERE DO THE FINDINGS COME FROM?

All findings in this briefing are extracted from the systematic review conducted by Ana Silva et al. (2017) titled "A systematic review on the use of Definition of Done on agile software development projects," presented at the Evaluation and Assessment in Software Engineering conference.

WHAT IS INCLUDED IN THIS BRIEFING?

The main findings of the systematic review, including insights into the criteria used for DoD and implications for practice.

WHAT IS NOT INCLUDED IN THIS BRIEFING?

Additional information not presented in the original systematic review, such as detailed descriptions of the individual studies analyzed.

For additional information about this briefing, please refer to the original paper: Silva, A., Araújo, T., Nunes, J., Perkusich, M., Dilorenzo, E., Almeida, H., & Perkusich, A. (2017). A systematic review on the use of Definition of Done on agile software development projects. In Proceedings of EASE'17, Kalskrona, Sweden. DOI: 10.1145/3084226.3084262.

11. To what extent do you agree with the following statement: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Slightly Disagree	I neither agree nor disagree	Slightly Agree	Agree	Strongly Agree
The conclusions presented in the briefing are consistent with those in the original article.	⬭	⬭	⬭	⬭	⬭	⬭	⬭

12. Please justify your choice

_____

_____

_____

_____

_____

13. To what extent do you agree with the following statement: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Slightly Disagree	I neither agree nor disagree	Slightly Agree	Agree	Strongly Agree
The level of certainty expressed in the briefing matches the confidence level presented in the original article.	◯	◯	◯	◯	◯	◯	◯

14. Please justify your choice

_____

_____

_____

_____

_____

15. To what extent do you agree with the following statement: *

Mark only one oval per row.

	Strongly Disagree	Disagree	Slightly Disagree	I neither agree nor disagree	Slightly Agree	Agree	Strongly Agree
All claims made in the briefing are supported by information present in the original article.	⬭	⬭	⬭	⬭	⬭	⬭	⬭

16. Please justify your choice