

AI as a research assistant?

Prof. Mika Mäntylä
mika.mantyla@helsinki.fi
University of Helsinki

Desire State - Press button and get results!

- Can AI do my research?
- Suitable tasks can include:
 - Perform include exclude screening in SLR
 - Extract data in SLR
 - Perform qualitative coding of textual data

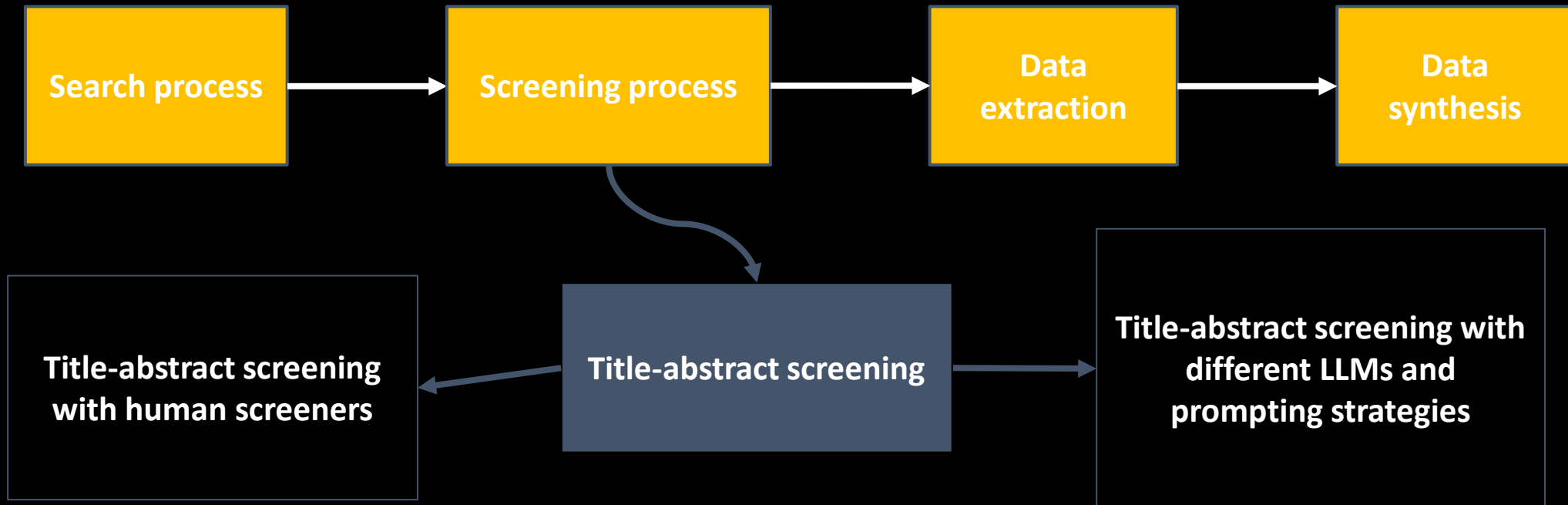


A. Huotala, M. Kuutila, P. Ralph, and M. Mäntylä, “*The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews*”. In Proceedings of the 28th (EASE '24)., 262–271

K. Romero Felizardo, M. Sampaio Lima, A. Deizepe, T. Uchôa Conte, and I. Steinmacher. *ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity*. In Proceedings of the ESEM '24.

Task 1: LLM for paper screening in SLR

Systematic review (SR)





Using LLMs to simplify abstracts improve human screening performance?

- Human subject experiment.
 - Human = post-docs, doctoral researchers and master's students
- Simplified abstracts did not improve screening accuracy
- Less time spent screening the simplified abstracts
- Non master's student status and TOSLS score predict accuracy

	Orig. abstract	Simp. abstract
Accuracy	69 %	69 %
Per-paper screening time	85.95 seconds	72.26 seconds

Comparison	p-value
Orig. vs Simp. Correctness	0.61
Orig. vs Simp. Time	0.0007***
Orig. vs Simp. confidence	0.75

LLMs vs Humans – Paper screening

A. Huotala, M. Kuutila, P. Ralph, and M. Mäntylä, “The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews”. In Proceedings of the 28th (EASE '24)., 262–271

- LLMs do not screen articles better than humans
- Several LLM & prompt combinations perform roughly as well as humans
- GPT4 – OS (one-shot) is better than average master’s students

LLM & prompt	coef	std.err	t	p-value	Odd’s ratio
Reference:	Human performance				
GPT3.5 - FS	-0.56	0.12	-4.56	<0.001	0.57
GPT3.5 - FS-CoT	-0.48	0.12	-3.86	<0.001	0.62
GPT3.5 - OS	-0.42	0.12	-3.38	<0.001	0.66
GPT3.5 - OSv2	-0.43	0.12	-3.44	<0.001	0.65
GPT3.5 - ZS	-0.91	0.12	-7.36	<0.001	0.40
GPT3.5 - ZSv2	-0.26	0.13	-2.12	0.03	0.77
GPT4 - FS	0.02	0.13	0.19	0.85	1.02
GPT4 - FS-CoT	0.13	0.13	1.03	0.30	1.14
GPT4 - OS	0.02	0.13	0.13	0.89	1.02
GPT4 - OSv2	-0.63	0.12	-5.08	<0.001	0.53
GPT4 - ZS	-0.20	0.13	-1.63	0.10	0.82
GPT4 - ZSv2	-0.31	0.12	-2.48	0.01	0.73
Reference:	Exclude				
Include	0.40	0.05	8.39	<0.001	1.50
Reference:	Original abs				
Simplified abstract	-0.08	0.06	-1.61	0.11	0.93

LLMs screening papers on likert scale

- 1-7 Likert scale whether paper should be included
 - 1 - Strongly disagree, 2 - Disagree, 3 - Somewhat disagree, 4 - Neither agree nor disagree, 5 - Somewhat agree, 6 - Agree, and 7 - Strongly agree
- Threshold 5 or greater
 - Best accuracy
 - FN too high -> Loses of evidence! (FN=16)
 - Some rework (FP=17)
- Threshold 4 or greater
 - Still, loses a lot of evidence! (FN = 14)
 - More rework (FP=35)
- Lower Threshold does not help

=>5	Predicted Positive	Predicted Negative
Actual Positive	48 (TP)	16 (FN)
Actual Negative	17 (FP)	53 (TN)
Accuracy	75.3%	

=>4	Predicted Positive	Predicted Negative
Actual Positive	50 (TP)	14 (FN)
Actual Negative	35 (FP)	35 (TN)
Accuracy	63.4%	

K. Romero Felizardo, M. Sampaio Lima, A. Deizepe, T. Uchôa Conte, and I. Steinmacher. *ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity*. In Proceedings of the ESEM '24.

Work in progress - Custom UI on top of foundation models (LLM)

AI-automated title-abstract screening PoC

Introduction

This proof-of-concept (PoC) demonstrates the capabilities of AI-automated title-abstract screening of systematic reviews (SRs), which is subject to further research and improvements. This PoC is based on a conference paper "The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews" by Huotala et al.

Supported LLMs

Currently, **we only support OpenAI models**. Support for additional providers is coming soon - via Openrouter.

License

CC-BY-ND 4.0

Source code

Available by request.

Examples

Load example

Step 1. Define screening criteria

Write a single criteria and press [Enter]

1. The main focus is time pressure in software engineering Delete
2. The paper presents empirical evidence of time pressure in software engineering Delete

Step 2. Add Title and Abstract

Time pressure in software engineering: A systematic review

Context

Large project overruns and overtime work have been reported in the software industry, resulting in additional expense for companies and personal issues for developers. Experiments and case studies have investigated the relationship between time pressure and software quality and productivity.

Objective

The present work aims to provide an overview of studies related to time pressure in software engineering; specifically, existing definitions, possible causes, and metrics relevant to time pressure were collected, and a mapping of the studies to software processes and approaches was performed. Moreover, we synthesize results of existing quantitative studies on the effects of

Step 3. OpenAI API key

OpenAI API key

Step 4. LLM configuration

Model

Load models

Temperature (0)

Seed

128

Enforce JSON output

Input token count

470

Estimated cost

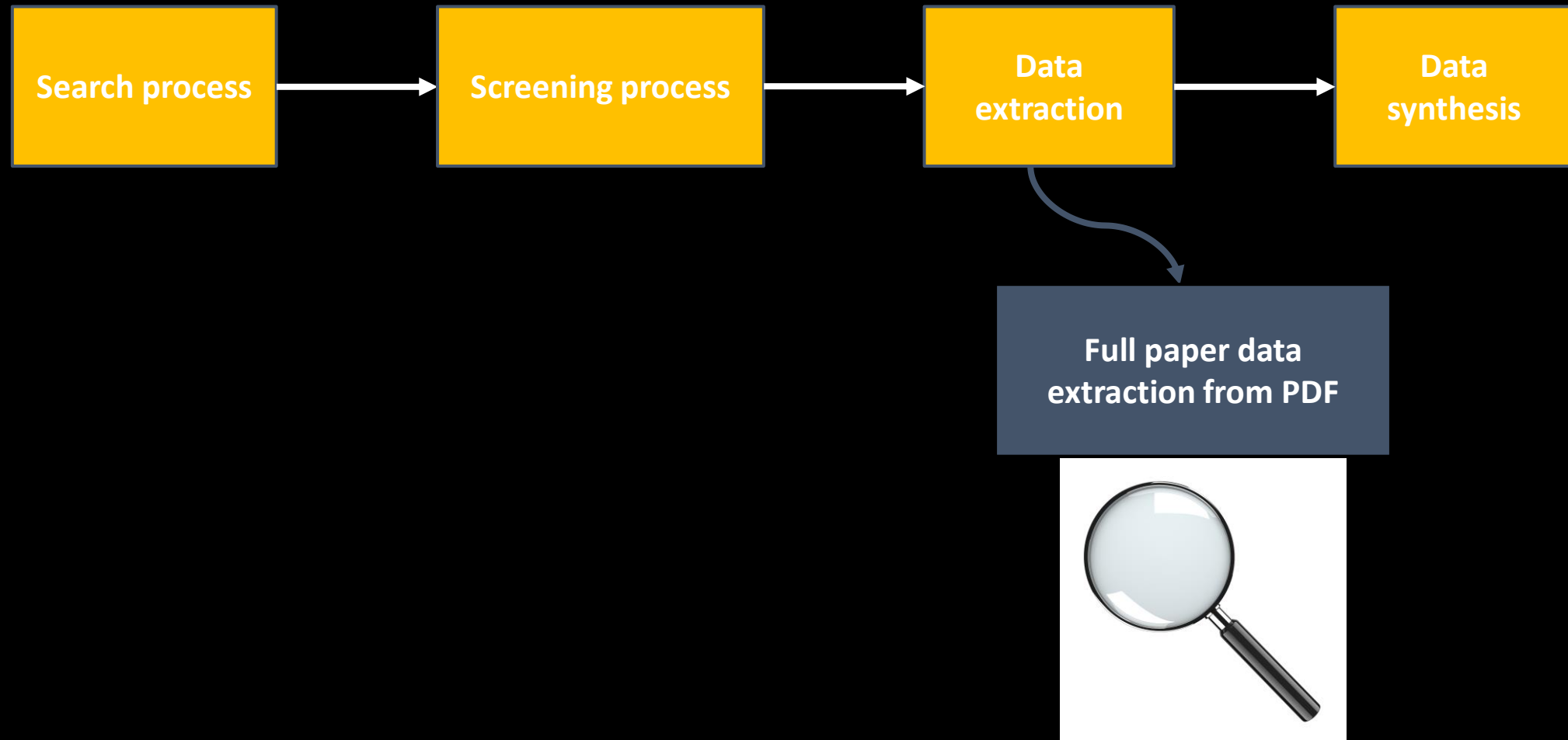
0.00148 \$

Send request

K. Romero Felizardo, I. Steinmacher, M. Sampaio Lima, Anderson Deizepe, T. Uchôa Conte, and M. Perini Barcellos. 2024. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In Proceedings of the (ESEM '24). 407–413.

Task 2: LLM for Systematic Mapping Data Extraction

Systematic review (SR)





ChatGPT application in SLRs in SE: an evaluation of its accuracy to support the selection activity



Katia Felizardo



Igor Steinmacher



Marcia Lima



Anderson Deizepe



Tayana Conte



Monalessa Barcellos



Igor.Steinmacher@nau.edu



LLM to extract systematic mapping data to JSON

Consider the following JSON structure:

```
[ {  
  "id": "",  
  "title": "",  
  "author1": "", "author2": "", "author3": "", "author4": "",  
  "author5": "", "author6": "", "author7": "", "author8": "",  
  "year": "",  
  "source_type": "",  
  "source_name": "",  
  "source_acronym": "",  
  
  "context": "",  
  "type_application": "", } Context  
  
  "strategy": "",  
  "instrument": "",  
  "data_source": null,  
  "psychological_mediators": null,  
  "behavioral_mediators": "" }  
]
```

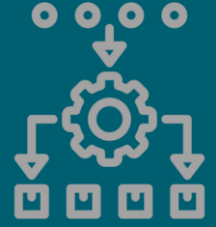
Meta-data

RQs

SLR: “A Journey to Identify Users’ Classification Strategies to Customize Game-Based and Gamified Learning Environments”

K. Romero Felizardo, I. Steinmacher, M. Sampaio Lima, Anderson Deizepe, T. Uchôa Conte, and M. Perini Barcellos. 2024. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In Proceedings of the (ESEM '24). 407–413.

Example of the data extracted



```
"id": "1",  
"title": "User Experience Design Using Machine Learning: A Systematic Review",  
"author1": "A.M.H. Abbas",  
"author2": "K.I. Ghauth",  
"author3": "C.-Y. Ting",  
"year": "2022",  
"source_type": "Journal",
```

Bibliometric data
(Objective data)

```
"type_application": "Machine Learning",  
"strategy": "Classification Algorithms",  
"instrument": "Surveys and Interviews",  
"data_source": "User Interaction Data",  
"psychological_mediators": "User Satisfaction",  
...
```

RQ's data

- › ChatGPT model: **ChatGPT – 4o**
- › Context window: **128,000 tokens** (to support the full text)



K. Romero Felizardo, I. Steinmacher, M. Sampaio Lima, Anderson Deizepe, T. Uchôa Conte, and M. Perini Barcellos. 2024. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In Proceedings of the (ESEM '24). 407–413.

SLR: “A Journey to Identify Users’ Classification Strategies to Customize Game-Based and Gamified Learning Environments”

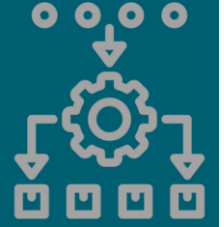
Results



	TP	TN	Misidenti- fied (FN)	Missed (FN)	Fabricated (FP)	Accuracy	Precision	Recall
Bibliometric data	213	0	3	0	0	98.61%	100%	98.61%
Context data	22	0	3	0	0	88.00%	100%	88.00%

K. Romero Felizardo, I. Steinmacher, M. Sampaio Lima, Anderson Deizepe, T. Uchôa Conte, and M. Perini Barcellos. 2024. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In Proceedings of the (ESEM '24). 407–413.

Research Method

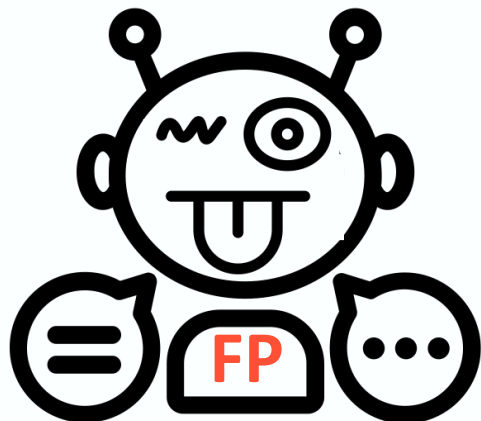


TP ⇒ number of data items ChatGPT correctly extracted

TN ⇒ number of data items ChatGPT correctly identified as unavailable in the full-text

FP ⇒ data items ChatGPT fabricated in cases where no data was available in the full-text

FN ⇒ number of *missing* data items or *misidentified* data items



Results



	TP	TN	Misidenti- fied (FN)	Missed (FN)	Fabricated (FP)	Accuracy	Precision	Recall
Bibliometric data	213	0	3	0	0	98.61%	100%	98.61%
Context data	22	0	3	0	0	88.00%	100%	88.00%
RQ1 (strategy)	20	0	6	0	0	76.92%	100%	76.92%
RQ2 (instrument)	20	0	3	0	2	80.00%	90.91%	86.96%

K. Romero Felizardo, I. Steinmacher, M. Sampaio Lima, Anderson Deizepe, T. Uchôa Conte, and M. Perini Barcellos. 2024. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In Proceedings of the (ESEM '24). 407–413.

Results

K. Romero Felizardo, I. Steinmacher, M. Sampaio Lima, Anderson Deizepe, T. Uchôa Conte, and M. Perini Barcellos. 2024. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In Proceedings of the (ESEM '24). 407–413.



	TP	TN	Misidentified (FN)	Missed (FN)	Fabricated (FP)	Accuracy	Precision	Recall
Bibliometric data	213	0	3	0	0	98.61%	100%	98.61%
Context data	22	0	3	0	0	88.00%	100%	88.00%
RQ1 (strategy)	20	0	6	0	0	76.92%	100%	76.92%
RQ2 (instrument)	20	0	3	0	2	80.00%	90.91%	86.96%
RQ3: data source	20	0	5	0	1	76.92%	95.24%	80.00%
RQ3: psychol. mediators	8	0	1	7	10	30.77%	44.44%	50.00%
RQ3: behav. mediators	12	4	0	1	9	61.54%	57.14%	92.31%
Total	321	4	15	8	22	87.84%	93.59%	93.31%

Results



	TP	TN	Misidenti- fied (FN)	Missed (FN)	Fabricated (FP)	Accuracy	Precision	Recall
Bibliometric data	213	0	3	0	0	98.61%	100%	98.61%
Context data	22	0	3	0	0	88.00%	100%	88.00%
RQ1 (strategy)	20	0	6	0	0	76.92%	100%	76.92%
RQ2 (instrument)	20	0	3	0	2	80.00%	90.91%	86.96%
RQ3: data source	20	0	5	0	1	76.92%	95.24%	80.00%
RQ3: psychol. mediators	8	0	1	7	10	30.77%	44.44%	50.00%
RQ3: behav. mediators	12	4	0	1	9	61.54%	57.14%	92.31%
Total	321	4	15	8	22	87.84%	93.59%	93.31%

Rantala, Leevi., Shar, Lwin Khin., & Mäntylä, Mika.
(2024). “*Studying SATD in DroneSystems with Human-AI Collaboration*” Submitted to a Journal, 1-32.

Task 3: LLM for Qualitative Data Coding

Study context

- Task classify self-admitted technical debt (SATD).
- SATD existing in source code comments
 - “Nasty hack to expose controller to unittest code”
 - “TODO Apparently can’t remove this or the build fails.”
- Classify: Architectural, Build, Code, Defect, Design, Doc, No Debt, Requirement, Test, Unclassified
- Data: 4,083 SATD comments
- 1 Human and 1 LLM (ChatGPT3.5) to classify

Results – Classifying 4,083 code comments

- Round 1 classification – Fleiss Kappa: 0.388 (Fair agreement) [-1, 1]
- Round 2 – Both annotators check disagreements
 - Two options are provided. One chosen initially and one chosen by the other
- Round 2
 - CGPT 3.5 changed its opinion: 1,481
 - Human changed its mind: 44
- Round 2 classification – Fleiss Kappa 0.83 (almost perfect)

Agreement & Disagreements

	Arc.	Build	Code	Defect	Design	Doc.	No Debt	Req.	Test	Unc.
Human/CGPT										
Architectural	72	8	11	1	5	1	1	8	0	15
Build	0	63	6	2	1	0	0	0	1	2
Code	3	6	1253	18	80	2	11	27	1	242
Defect	1	2	16	459	12	2	1	14	0	51
Design	3	3	58	13	426	1	1	60	0	138
Documentation	0	0	1	1	0	55	0	1	0	0
No Debt	0	0	0	0	0	0	0	0	0	0
Requirement	4	2	13	12	54	5	0	466	0	96
Test	0	1	4	2	0	1	2	5	97	8
Unclassifiable	1	2	4	4	5	0	0	3	0	747

Rantala, Leevi., Shar, Lwin Khin., & Mäntylä, Mika. (2024). “*Studying SATD in DroneSystems with Human-AI Collaboration*” Submitted to a Journal, 1-32.

Agreement & Disagreements

	Arc.	Build	Code	Defect	Design	Doc.	No Debt	Req.	Test	Unc.
Human/CGPT										
Architectural	72	8	11	1	5	1	1	8	0	15
Build	0	63	6	2	1	0	0	0	1	2
Code	3	6	1253	18	80	2	11	27	1	242
Defect	1	2	16	459	12	2	1	14	0	51
Design	3	3	58	13	426	1	1	60	0	138
Documentation	0	0	1	1	0	55	0	1	0	0
No Debt	0	0	0	0	0	0	0	0	0	0
Requirement	4	2	13	12	54	5	0	466	0	96
Test	0	1	4	2	0	1	2	5	97	8
Unclassifiable	1	2	4	4	5	0	0	3	0	747

77.5% agreement overall
 CGPT says Unclassified: 57.5% agreement
 If we remove Unclassified we get 85.1% overall agreement

Rantala, Leevi., Shar, Lwin Khin., & Mäntylä, Mika. (2024). “*Studying SATD in DroneSystems with Human-AI Collaboration*” Submitted to a Journal, 1-32.

Agreement & Disagreements

Human/CGPT	Arc.	Build	Code	Defect	Design	Doc.	No Debt	Req.	Test	Unc.
Architectural	72	8	11	1	5	1	1	8	0	15
Build	0	63	6	2	1	0	0	0	1	2
Code	3	6	1253	18	80	2	11	27	1	242
Defect	1	2	16	459	12	2	1	14	0	51
Design	3	3	58	13	426	1	1	60	0	138
Documentation	0	0	1	1	0	55	0	1	0	0
No Debt	0	0	0	0	0	0	0	0	0	0
Requirement	4	2	13	12	54	5	0	466	0	96
Test	0	1	4	2	0	1	2	5	97	8
Unclassifiable	1	2	4	4	5	0	0	3	0	747

Best: CGPT classifies as Test Debt -> Human agrees 98% of the time

Rantala, Leevi., Shar, Lwin Khin., & Mäntylä, Mika. (2024). “*Studying SATD in DroneSystems with Human-AI Collaboration*” Submitted to a Journal, 1-32.

Agreement & Disagreements

Human/CGPT	Arc.	Build	Code	Defect	Design	Doc.	No Debt	Req.	Test	Unc.
Architectural	72	8	11	1	5	1	1	8	0	15
Build	0	63	6	2	1	0	0	0	1	2
Code	3	6	1253	18	80	2	11	27	1	242
Defect	1	2	16	459	12	2	1	14	0	51
Design	3	3	58	13	426	1	1	60	0	138
Documentation	0	0	1	1	0	55	0	1	0	0
No Debt	0	0	0	0	0	0	0	0	0	0
Requirement	4	2	13	12	54	5	0	466	0	96
Test	0	1	4	2	0	1	2	5	97	8
Unclassifiable	1	2	4	4	5	0	0	3	0	747

2nd Best: CGPT classifies as Code Debt -> Human agrees 92%

Summary

- Task 1 – LLMs in SLR screening (Title-abstract)
 - LLM simplified abstracts are quicker to screen for humans without penalty in accuracy
 - Screening accuracy: LLM => master's student
 - Screening accuracy: LLM < doctoral research / post-doc
 - Screening accuracy: LLM miss evidence
 - Attempts to fix this-> Most evidence still missed with big increase in false positives.
- Task 2 – LLM in SLR data extraction from PDF
 - Bibliometric, e.g. author name -> Accuracy 98%
 - More difficult Psychological mediators -> Accuracy 35%
- Task 3 – LLM in qualitative coding (classification)
 - Can act as second data coder
 - Two rounds and good agreement even with CGPT 3.5
 - Human had made 40 errors (1% of data)
 - Correct classification rates vary 58% to 98%

A. Huotala, M. Kuutila, P. Ralph, and M. Mäntylä, "The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews". In Proceedings of the 28th (EASE '24)., 262–271

K. Romero Felizardo, M. Sampaio Lima, A. Deizepe, T. Uchôa Conte, and I. Steinmacher. *ChatGPT application in Systematic Literature Reviews in Software Engineering: an evaluation of its accuracy to support the selection activity*. In Proceedings of the ESEM '24.

K. Romero Felizardo, I. Steinmacher, M. Sampaio Lima, Anderson Deizepe, T. Uchôa Conte, and M. Perini Barcellos. 2024. Data extraction for systematic mapping study using a large language model - a proof-of-concept study in software engineering. In Proceedings of the (ESEM '24). 407–413.

Rantala, Leevi., Shar, Lwin Khin., & Mäntylä, Mika. (2023). "Studying SATD in Drone Systems with Human-AI Collaboration" Submitted to a Journal, 1-32.

Conclusion

Press button and get results?

- Can AI do my research? Not yet
- You are still the captain
 - LLM is co-pilot only
- LLMs can get lot of things correct 80-95%.
- You can use LLMs only if you are the expert!
 - Otherwise, one cannot recognize incorrect 5-20%

