



Case study identification with GPT-4 and implications for mapping studies

Kai Petersen*

Blekinge Institute of Technology, Valhallavägen 1, 37141 Karlskrona, Sweden

Flensburg University of Applied Sciences (FUAS), Kanzleistraße 91-93, 24943 Flensburg, Germany

ARTICLE INFO

Keywords:

Systematic mapping studies
Data extraction
Case study
GPT-4

ABSTRACT

Context: Rainer and Wohlin showed that case studies are not well understood by reviewers and authors and thus they say that a given research is a case study when it is not.

Objective: Rainer and Wohlin proposed a smell indicator (inspired by code smells) to identify case studies based on the frequency of occurrences of words, which performed better than human classifiers. With the emergence of ChatGPT, we evaluate ChatGPT to assess its performance in accurately identifying case studies. We also reflect on the results' implications for mapping studies, specifically data extraction.

Method: We used ChatGPT with the model GPT-4 to identify case studies and compared the result with the smell indicator for precision, recall, and accuracy.

Results: GPT-4 and the smell indicator perform similarly, with GPT-4 performing slightly better in some instances and the smell indicator (SI) in others. The advantage of GPT-4 is that it is based on the definition of case studies and provides traceability on how it reaches its conclusions.

Conclusion: As GPT-4 performed well on the task and provides traceability, we should use and, with that, evaluate it on data extraction tasks, supporting us as authors.

1. Introduction

When conducting qualitative studies, we rely on text interpretation to extract information. We need to determine different properties of primary studies in literature studies, just as identifying the research type or research methods correctly (see e.g., [1]). In earlier work, Rainer and Wohlin [2] identified issues when authors of primary studies and systematic review authors label studies as case studies. Wrongly labeling case studies is problematic as the characteristics of case studies (deep insights through triangulation and studying real world insights) could be inferred from the data. Thus, aggregated results could be misleading.

In response, they proposed a simple indicator to determine whether a study is a case study. The occurrences of word interview used as a case study predictor. The indicator performed very well, achieving accuracy values from 71% to 88% on different data sets. They define a case study as an empirical investigation of a case, using multiple data collection methods, to study a contemporary phenomenon in its real-life context, and with the investigator(s) not taking an active role in the case investigated. The definition was derived from previous guidelines on case study research, such as Yin [3] and Runeson and Höst [4]. The authors acknowledge one major drawback, namely that their indicator is not associated with the characteristics of case studies. That is, their

indicator is purely based on the occurrence of terms, which does not weigh all case study criteria into the assessment.

With the emergence of large language models (LLMs), such as GPT-3 and GPT-4, new possibilities for text analysis emerged. ChatGPT also provides various tools that facilitate the analysis of academic texts. ChatGPT is accessed through prompts that describe a task. The resolution of the task is sent back as a response.

This study aims to evaluate GPT-4 in its ability to detect case studies in a traceable way (i.e. GPT-4 documenting how it reached its conclusion). We make the following contributions:

1. Develop and improve a prompt for case study identification.
2. Assess the prompt against three data sets used by Rainer and Wohlin [2].

We also reflect on the broader implications of GPT-4 in literature studies. The motivation to focus on case study identification was two-fold: (1) literature observed a problem of wrongly labeling case studies; (2) labeled data sets are available in [2].

2. Method

We aimed to answer the following research questions:

* Correspondence to: Blekinge Institute of Technology, Valhallavägen 1, 37141 Karlskrona, Sweden.

E-mail addresses: kai.petersen@hs-flensburg.de, kai.petersen@bth.se.

- RQ1: How to formulate a prompt for ChatGPT to identify case studies in software engineering accurately?
- RQ2: How do the indicators perform in comparison to the measures used by Rainer and Wohlin?

GPT-4 is evaluated using the same data utilized in Rainer and Wohlin’s study [2].

2.1. Data set

Three data sets were provided by Rainer and Wohlin, comprising Data Set 1 (68 studies), Data Set 2 (35) and Data Set 3 (100 articles), respectively. We could not retrieve one article due to missing subscriptions for Data Set 1. Therefore, we only analyzed the 67 available studies. Rainer and Wohlin (see [2] for details) characterize their data sets as follows.

1. Data Set 1: Stratified sample of 68 studies from nine Systematic Literature Reviews (SLRs) to achieve a broad perspective.
2. Data Set 2: Focused sample of 35 studies from one SLR.
3. Data Set 3: 100 Studies used in an assessment by [5] determining whether case studies were correctly labeled.

2.2. Data collection procedure

We utilized the definition of case studies presented in the paper by Rainer and Wohlin. We formulated a description for GPT-4 for each criterion. Data Set 1 was used with an initial prompt. For Data Set 2 and 3 the prompt was improved in between. The final prompt proposed is shown in Fig. 1.

As a reference for comparison, we relied on the outcome of the Smell Indicator provided in the original paper [2]. The smell indicator was rule-based and was defined as (quote): “given a primary study

already classified as a case study (precondition), then the primary study is indicated to be a case study if the term “interview” occurs at least three times in the article reporting the study”.

2.3. Analysis

We used accuracy as a measure that was used in [2]. In addition, we calculated precision and recall based on the data provided by [2]. We compared GPT-4 with the ground truth (assessed by [2,5]) and calculated all measures based on the data provided to compare with the smell indicator.

2.4. Study validity

Rainer and Wohlin highlighted that their judgment of case studies was not based on a systematic assessment of inter-rater reliability. As they acknowledge, this is a threat to the validity and reliability of the assessment. Given that we use GPT-4, exact replication is impossible, as the language model may sometimes provide different responses. As we also focused on the same set of studies, external validity concerns also apply, i.e., the results may not be transferable to other fields.

3. Results

Table 1 compares the proficiency of GPT-4 and the Smell Indicator in identifying case studies across three data sets. In Data Set 1, both approaches demonstrated comparable precision, with the Smell Indicator slightly outperforming GPT-4. However, GPT-4 had a higher recall rate, while the Smell Indicator boasted a marginally better accuracy. For Data Set 2, the Smell Indicator is better at capturing most of the true case studies, but when GPT-4 predicts something as a case study, it is more likely to be correct. In Data Set 3, the Smell Indicator showcased

ChatGPT Prompt
<p>A case study is an empirical investigation of a case, using multiple data collection methods, to study a contemporary phenomenon in its real-life context, and with the investigator(s) not taking an active role in the case investigated". A case study can be recognized by the following characteristics:</p> <ol style="list-style-type: none"> a) It describes a concrete case (such as a product, project, company, or other) that was investigated in the context of a study. b) It investigates a contemporary phenomenon, which means the phenomenon exists or occurs at the present time. c) It refers to real-life context, meaning a real project, product, or similar is studied (e.g. company, open source). Lab experiments with students or small self-constructed examples are not cases. d) It uses It uses multiple data collection methods. Data collection methods, among others, are: \textbf{interviews, focus groups, questionnaires, observations, document analysis, archival research, post-mortem review, think-aloud, and measurements of software artifacts and entities (e.g. project metrics, product metrics, test metrics, etc.)}. If only one method is mentioned, this is not fulfilled. e) The investigator does not take an active role in the case investigated. That is, the investigator collects data but does not influence or take an active role in the products or projects studied. If there is evidence of the investigator taking an active role, this should be assessed as not fulfilled. If no judgment is possible, e) should be regarded as fulfilled
ChatGPT Response
<p>Based on the information extracted from the document, let's assess the paper against the criteria for a case study:</p> <ol style="list-style-type: none"> a) Quote: "In the last stage, we examined data collected in a longitudinal multiple case study of two companies engaged in scaling agile development adoption according to the DAD (Case A) and SAFe (Case B) frameworks." [Page 8]; \textbf{Assessment: Fulfilled} b) Quote: "During our four year collaboration, authors 1 and 4, along with their colleagues, conducted 31 interviews of team members and managers in a variety of roles, at all levels of the company, chosen to be representative of all levels of the development organization." [Page 9]; \textbf{Assessment: Fulfilled} c) "The business model for both cases is to develop, maintain and sell software to clients throughout the globe." [Page 8]. \textbf{Assessment: Fulfilled} d) Quote: "During our four year collaboration, authors 1 and 4, along with their colleagues, conducted 31 interviews of team members and managers in a variety of roles, at all levels of the company... We also observed distributed development teams conducting Scrum 'ceremonies,' such as daily standups, sprint planning, and retrospective meetings; and, we observed weekly program-level 'scrum of scrums' style meetings." [Page 9]; \textbf{Assessment: Fulfilled} e) Quote: "We are indebted to our Case Study organizations—Company A and Ocuco Ltd (Company B)—and the many members of those organizations we observed and interviewed over a four-year period." [Page 29]; <p>Assessment: There is no direct evidence of the investigator taking an active role in the case. Based on the information provided, it seems the investigators observed and interviewed without influencing the case. Fulfilled.</p> <p>Overall Judgment: Given that all the criteria (a to f) are fulfilled, the study can be classified as a case study.</p> <p>Confidence Level: High</p>

Fig. 1. ChatGPT conversation.

Table 1
Comparison of measures for GPT-4 and smell indicator across data sets.

Data set	Predictor	Precision	Recall	Accuracy
Data Set 1 (Prompt v1)	GPT-4	0.348	0.727	0.731
	Smell Indicator	0.375	0.545	0.776
Data Set 2 (Prompt v2)	GPT-4	0.706	0.75	0.743
	Smell Indicator	0.609	0.875	0.686
Data Set 3 (Prompt v3)	GPT-4	0.746	0.830	0.76
	Smell Indicator	0.977	0.793	0.88

great precision at 0.977, higher than GPT-4's 0.746. However, GPT-4 had a higher recall rate. Overall, in terms of accuracy across all data sets, the Smell Indicator demonstrated comparable results to GPT-4.

4. Discussion

Use GPT to challenging your classifications: ChatGPT provides detailed feedback on how it reached the decisions, with traceability to text. We can use the responses to challenge the author's assessment. In Study 16 (Data Set 3), assessed as a case study by Wohlin [5], it detected that the study was action-research like, not fulfilling Criterion (d) of our assessment. Specifically, ChatGPT pointed out that the researchers took an active collaborative role in the case, citing a quote from the paper: "As Altinn's research partner, we applied principles from action research11 to guide our collaboration with Altinn: 1. The Principle of the Researcher-Client Agreement: To become involved in the case as a key source of learning, we took an active collaborative role". [Page 5]

Potential for further improvement: One potential we see is to preprocess the data set, chunking the papers into sections and providing the relevant sections for an extraction item (e.g., the method and results section for extracting information about the research method). We also see a potential in repeatedly running the extraction and assessing the intra-rater reliability as the process of analysis is non-deterministic. Thereby, we could simulate multiple authors doing the task.

Implications for mapping studies: GPT may contribute to mapping studies in the following ways:

- Use the language model results in combination with the Smell indicator for case study assessment. As they could extract case study-related information, we may use them to conduct other extractions as well (e.g., what research type is used, such as evaluation or validation research). GPT-4 and SI could thus be used to triangulate the reviewers' decisions.
- Reflect on the output of the model and contrast your own extraction against the one done by the model. Using GPT adds the following benefits: (1) we receive a written assessment and hence have traceability how the decision came about, with a direct connection to the definition of what case studies are. (2) we have

traceability to phrases used in the assessed papers, that allow us to control and check the judgment made by GPT. Overall, this makes GPT a good "sparring partner" during data extraction to confirm or challenge our own assessments.

5. Conclusion

In this study, we used GPT-4 to identify case studies. We ran the study on three data sets provided in an earlier study suggesting a Smell Indicator for case study identification. Overall, we saw that GPT-4 provides similar quantitative results concerning accuracy, precision and recall. The added value of GPT-4 is the capability of explaining how it reached its conclusions. GPT advances simple term frequency analysis by utilizing embeddings and neural networks within the transformer architecture. This advancement, along with an increased context window in newer models, enhances GPT's ability to perform data extraction and interpret complex scenarios beyond basic term associations.

CRedit authorship contribution statement

Kai Petersen: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.infsof.2024.107452>.

References

- [1] Kai Petersen, Robert Feldt, Shahid Mujtaba, Michael Mattsson, Systematic mapping studies in software engineering, in: 12th International Conference on Evaluation and Assessment in Software Engineering, EASE 12, 2008, pp. 1–10.
- [2] Austen Rainer, Claes Wohlin, Case study identification: A trivial indicator outperforms human classifiers, *Inf. Softw. Technol.* (2023) 107252.
- [3] Robert K. Yin, *Case Study Research: Design and Methods*, vol. 5, Sage, 2009.
- [4] Per Runeson, Martin Höst, Guidelines for conducting and reporting case study research in software engineering, *Empir. Softw. Eng.* 14 (2009) 131–164.
- [5] Claes Wohlin, Case study research in software engineering—It is a case, and it is a study, but is it a case study? *Inf. Softw. Technol.* 133 (2021) 106514.