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), Kanesstræde 91-93, 24943 Flensburg, Germany

A B S T R A C T

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 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:
• 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...
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**

[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.