T5- only a small subset of production systems is studied, limiting the impact of this research . performing replications requires a deep understanding of research methodologies and subtle nuances in software engineering data . large language models (LLMs) such as GPT-4 show promise in tackling both software engineering- and science-related tasks . their ability to surface assumptions made in empirical software engineering research methodologies .

Bart -Can GPT-4 Replicate Empirical Software Engineering Research?. We find that G PT-4 is able to surface correct assumptions but struggles to generate ones that apply common knowledge about software engineering data. Our findings have implications for leveraging LLMs for software engineering research as well as practitioner data scientists in software teams. In a manual analysis of the generated code, we find that the GPT.-4-generated code contains correct high-level logic, given a subset of the methodology. However, the code contains many small implementation-level errors, reflecting a lack of software engineering knowledge. We perform a user study with 14 participants with software. engineering research expertise, who evaluate GPT.-4- generated assumptions and analysis plans (i.e., a list of module specifications) from the papers. We found that Gpt-4 was able tosurface correct assumptions, but struggled to generate those that applycommon knowledge aboutSoftware engineering data, such as the use of the word “engine’s” in the name of a module.

Gemini - GPT-4 can assist in replicating empirical software engineering research, enabling practitioners to benefit from research findings on their own data. However, GPT-4 faces challenges in identifying assumptions about software engineering data and generating error-free code for analysis pipelines. Despite these limitations, GPT-4 has potential to democratize empirical software engineering research by reducing the expertise required for replications.

Gpt-4 - The research paper explores the potential of GPT-4, a large language model (LLM), in replicating empirical software engineering research, a field that significantly enhances understanding but is often limited by the narrow range of production systems studied. This limitation restricts the broader applicability of research findings, making it challenging for practitioners to replicate studies on their own data due to the complex nuances and deep understanding required of research methodologies and software engineering data.

The study focuses on GPT-4's capabilities to assist in this replication process by examining its ability to identify underlying assumptions in research methodologies and to develop code for analysis pipelines based on seven selected empirical software engineering papers. A user study involving 14 participants with expertise in software engineering research was conducted to evaluate the assumptions and analysis plans generated by GPT-4. The findings reveal that while GPT-4 successfully identifies correct assumptions, it struggles with incorporating common knowledge about software engineering data into these assumptions.

Further, the analysis of the code generated by GPT-4 shows that although the model manages to capture the correct high-level logic required by the research methodologies, it frequently makes small, implementation-level errors. These errors indicate a gap in the model's deep understanding of software engineering practices, which could potentially hinder its utility in practical applications.

The implications of these findings are significant for both the field of software engineering research and for data scientists working within software teams. They suggest that while LLMs like GPT-4 hold promise for democratizing and expanding the reach of empirical software engineering research, there are still challenges to overcome, particularly in terms of the model's understanding of complex software engineering concepts and data. This research underscores the need for further development and refinement of LLMs to better support software engineering research and practice.

Gpt-3 -Empirical software engineering research has enhanced the understanding of software engineering processes, yet its impact is limited due to the small subset of production systems studied. Replicating research on their own data could benefit practitioners, but challenges arise due to the complexities of research methodologies and nuances in software engineering data. Large language models like GPT-4 offer promise in democratizing this research by aiding in replications. This study evaluates GPT-4's ability to replicate empirical software engineering research by analyzing assumptions and generating code for analysis pipelines from seven papers. A user study with 14 software engineering experts reveals that GPT-4 can identify correct assumptions but struggles with generating those based on common software engineering knowledge. While the generated code demonstrates accurate high-level logic aligned with the methodology, it contains numerous implementation-level errors due to a lack of software engineering expertise. These findings highlight the potential of leveraging LLMs for software engineering research and the importance of integrating software engineering knowledge in data science practices within software teams. Ultimately, this research sheds light on the opportunities and challenges associated with using advanced language models like GPT-4 in replicating and advancing empirical software engineering research.