Limitations of ChatGPT in Conceptual Modeling: Insights from Experiments in Metamodeling

Fabian Muff [1] and Hans-Georg Fill [1]

Keywords: LLM, Conceptual Modeling, Metamodeling

Extended Abstract

Recent years have seen significant progress in machine learning technology, leading to the development of large language models (LLMs) like ChatGPT[2] and Bard[3], which are currently being investigated in various fields. LLMs already play a role in conceptual modeling research [1, 2]. In this context, we describe insights we gained from experiments for analyzing metamodels using large-language models. The goal of the experiments was to assess to what extent large language models such as used by ChatGPT-4 are able to aid in the processing of state-of-the-art metamodels. In this context, we were particularly interested in whether an LLM could sufficiently understand a complex language definition as used in conceptual modeling tools, and what limitations it would face.

Method and Experiments

Our experimental research approach involved the following explorations of metamodels in JSON format, which is created by our web-based metamodeling platform MM-AR [3]. We provide the JSON files in an online repository[4].

Analysis of Metamodels: The experiments began with an analysis of ten different metamodels using ChatGPT. These metamodels varied in size and complexity. The main objective was to determine whether ChatGPT could correctly analyze the structure and semantic constraints.
Generation of Model Transformations: Next, we assign two primary objectives to ChatGPT. First, to analyze the metamodels and then generate a TypeScript program to transform a metamodel into a PlantUML[5] class diagram. This was based on two inputs to the LLM: a natural language description of the meta-metamodel plus the actual metamodel in JSON

1 University of Fribourg, Fribourg, Switzerland, fabian.muff@unifr.ch, https://orcid.org/0000-0002-7283-6603; hans-georg.fill@unifr.ch, https://orcid.org/0000-0001-5076-5341
2 https://openai.com/blog/chatgpt
3 https://bard.google.com/chat
4 https://doi.org/10.5281/zenodo.10695823
5 https://plantuml.com/en/guide

notation. The deterministic nature of the TypeScript program provided us with a baseline to compare it to the PlantUML code directly generated by the LLM.
Model Instance Generation: The last experiment encompassed the creation of model instances from a given metamodel for the Business Process Model and Notation (BPMN). This included again a description of the meta-metamodel, the BPMN metamodel in JSON notation, a sample BPMN model instance in JSON, and a natural language description of a business process which should be represented by the model. This step aimed to evaluate the LLMs ability to handle the instantiation of metamodels.

Findings

The experiments conducted with ChatGPT 4.0 yielded interesting insights and revealed the limitations of LLMs at their current stage. In the first experiment, ChatGPT was unable to process the metamodels in their original size (757 KB of code). Rather, we had to remove certain information, such as the contained graphical representations, and reduce the size to approximately 18 KB of code. For the interpretation of the metamodels we had to add natural-language descriptions of the structure of the JSON file and the relationships to retrieve useful output. Secondly, the generated model transformation scripts were highly effective, requiring only minimal adjustments primarily of file paths and names. Accurate transformations into PlantUML code were created for ten different metamodels. However, generating PlantUML class diagrams directly by the LLM posed significant challenges. ChatGPT 4.0 struggled due to the complexity and length of the PlantUML code, especially for extensive metamodels with various combinations of class relations. Additionally, the outputs of this task were highly variable due to its nondeterministic nature. Even for identical inputs, the generated PlantUML scripts lacked consistency and predictability. Thirdly, despite many adjustments and additional explanations in natural language, we did not succeed in generating correct model instances from a given metamodel with ChatGPT.

Conclusion

In summary, the experiments with ChatGPT 4.0 confirmed its potential in standard programming tasks. However, it highlighted the current limitations of the LLM in understanding and handling complex data structures. We are confident that these findings provide valuable insights for the further application of LLMs in conceptual modeling.

References

[1] Hans-Georg Fill, Peter Fettke, and Julius Köpke. “Conceptual Modeling and Large Language Models: Impressions From First Experiments With ChatGPT”. In: *Enterp. Model. Inf. Syst. Archit. Int. J. Concept. Model.* 18 (2023), p. 3. DOI: 10.18417/EMISA.18.3. URL: https://doi.org/10.18417/emisa.18.3.

[2] Hans-Georg Fill and Fabian Muff. “Visualization in the Era of Artificial Intelligence: Experiments for Creating Structural Visualizations by Prompting Large Language Models”. In: *CoRR* abs/2305.03380 (2023). DOI: 10.48550/arXiv.2305.03380.

[3] Fabian Muff and Hans-Georg Fill. “Initial Concepts for Augmented and Virtual Reality-based Enterprise Modeling”. In: *ER Demos and Posters 2021*. Vol. 2958. CEUR-WS.org, 2021, pp. 49–54. URL: https://ceur-ws.org/Vol-2958/paper9.pdf.