GPT4 aided biomaterials research use case: stabilization of selenium nanoparticles with proteins

Zoran Stojanović*, Nenad Filipović, Maja Kuzmanović, Sara Lukač and Magdalena Stevanović

Institute of Technical Sciences of SASA, Belgrade, Republic of Serbia

Keywords: Material science; chemistry; prompt-engineering; artificial intelligence

Hem. Ind. **78(1S)** 68 (2024)

Available on-line at the Journal web address: http://www.ache.org.rs/HI/

Recent advancements in LLMs based on various transformer architectures such as BERT and GPT family models [1], brought many new possibilities for application in scientific research. The specific architecture and broad knowledge of these models give them the ability to understand concepts, to plan and solve different kinds of problems, including various chemistry-related tasks [2,3]. In this work, we are evaluating a case of GPT4 performance for recommending proteins suitable for the stabilization of selenium nanoparticles (SeNPs). SeNPs exhibit diverse beneficial bioactivities, including antioxidant, antibacterial, and anticancer properties, and stabilization of SeNPs with suitable proteins may be an effective approach to improve their bioactivities.

Initially, we made a series of "zero-shot" tests to evaluate knowledge, problem-solving ability, and identify weaknesses of GPT4 on the research topic and subsequently, we optimized prompts if needed to get correct responses. Test questions in related domains of crystallography, colloidal chemistry, and biochemistry were selected from free sources online (e.g. sanfoundry.com) or created by us. Finally, we have evaluated model performances on the main task which is to suggest the best protein candidates for stabilization of SeNPs that we synthesized in laboratory experiments. The design of prompts was done according to proposed tactics for prompt engineering suggested by OpenAI and DAIR.AI [4]. For interaction with GPT4 model OpenAI, API and Python programming language were employed.

The model successfully completed all our benchmark tests with optimized prompts. It also demonstrated the ability to analyze procedures for the synthesis and stabilization of SeNPs. GPT4 exhibits the capability to recognize protein records by identifiers, correctly identify amino acid sequences, describe their properties and functions to some extent depending on data trained on, identify available binding sites for interaction with Se, and give, by our judgment, good proposals for SeNPs stabilizers.

The study demonstrates the successful application of advanced transformer architecture models like GPT4 in addressing relatively complex tasks in materials research. Despite GPT4 capabilities being largely dependent on the quality and size of training data, utilization of strategically designed and optimized prompts significantly improves it's performance.

REFERENCES

[1] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, X. Hu, Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, arXiv:2304.13712 27 April 2023, https://doi.org/10.48550/arXiv.2304.13712

[2] A. D. White, G. M. Hocky, H. A. Gandhi, M. Ansari, S. Cox, G. P. Wellawatte, S. Sasmal, Z. Yang, K. Liu, Y. Singh, W. J. Peña Ccoa, Assessment of chemistry knowledge in large language models that generate code. *ChemRxiv 2022* https://doi.org/10.26434/chemrxiv-2022-3md3n-v2

[3] K. Hatakeyama-Sato, N. Yamane, Y. Igarashi, Y. Nabae & T. Hayakawa, Prompt engineering of GPT-4 for chemical research: what can/cannot be done?, STAM Methods, 3:1, 2260300, 2023, https://doi.org/10.1080/27660400.2023.2260300

[4] https://www.promptingguide.ai/

*Corresponding author E-mail: zoran.stojanovic@itn.sanu.ac.rs