Knowledge-driven Scientific Large Language Models

Qiang Zhang

Zhejiang University - University of Illinois Urbana-Champaign Institute, International Campus, Zhejiang University
718 Haizhou East Road, Haining 314400, China
qiang.zhang.cs@zju.edu.cn

Humanity acquires knowledge of the world via perception and cognition, where *natural languages* (i.e., human languages) stand as the quintessential medium for articulating this *world knowledge*. Historically, this plethora of world knowledge has been expressed, chronicled, and disseminated through natural languages. Currently, *Large Language Models* (LLMs) stand as cutting-edge tools in processing natural language and gathering world knowledge.

Besides natural languages, to encapsulate more specialized *science knowledge*, an assortment of *scientific languages* has been developed. This encompasses textual expressions in the scientific research domains, mathematical languages to define mathematical formulas, chemical languages such as SMILES that represent molecular structures, and biological languages that describe proteins or genomes, and detail the complex constitution of living organisms. These scientific languages come with their distinct vocabularies, where each term holds a specific meaning that can be completely different from natural languages. Furthermore, experts in specific domains establish grammatical rules to organize these terms, enabling the construction of sentences with precise semantic functions. Due to the potential semantic and grammatical differences between scientific and natural languages, existing general LLMs often fail to properly deal with scientific data like molecules and proteins.

To facilitate the understanding of scientific languages, researchers have devised *Scientific Large Language Models* (Sci-LLMs) for various domains. For instance, molecular language models have been developed to represent molecule structures as a string of atoms and chemical bonds. These models aid in predicting molecular properties, designing new drugs, and proposing retrosynthesis routes. Similarly, protein language models operate based on sequences of amino acids. They are used to forecast 3D protein structures and functions, enhance existing proteins for improved fitness, and create new proteins with specific functionalities.

In this talk, I will summarize my research on Sci-LLMs with close reference to general LLM advancements. Given the broad scope of scientific languages, I will focus particularly on biological and chemical languages. Specifically, my investigation will encompass the following areas:

- **Scientific Large Language Models:** My research explores the development and application of Sci-LLMs for various scientific domains, including textual, molecular, protein, and genomic languages. These models, such as *InstructProtein* (Wang et al. 2024), are designed to understand, generate, and align multiple types of scientific languages, enabling more efficient protein design, molecule generation, and functional prediction.
- **Knowledge Graph Integration:** A key focus of my research is the integration of knowledge graphs (KGs) with scientific LLMs. KGs offer a structured way to represent complex relationships between entities in scientific domains. In models like the *KANO* (Fang et al. 2023), I demonstrated how KG-enhanced pre-trained models can improve molecular predictions, including drug discovery and retrosynthesis.
- **Sci-LLM Evaluation:** Another critical aspect of my work involves the development of comprehensive evaluation frameworks for Sci-LLMs. Through benchmarks like *SciKnowEval* (Feng et al. 2024), I aim to measure the performance of LLMs across multiple scientific domains, ensuring that models meet the accuracy, safety, and ethical standards required for high-stakes applications such as drug design and environmental sustainability.

Through this exploration, I aim to provide insights into the intersection of LLMs with scientific languages, discussing both current advancements and future directions, such as safety, embodied AI, and the challenges in aligning scientific knowledge across different domains.

References

Fang, Y.; Zhang, Q.; Zhang, N.; Chen, Z.; Zhuang, X.; Shao, X.; Fan, X.; and Chen, H. 2023. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 1–12.

Feng, K.; Ding, K.; Wang, W.; Zhuang, X.; Wang, Z.; Qin, M.; Zhao, Y.; Yao, J.; Zhang, Q.; and Chen, H. 2024. SciKnowEval: Evaluating Multi-level Scientific Knowledge of Large Language Models. *arXiv preprint arXiv:2406.09098*.

Wang, Z.; Zhang, Q.; Ding, K.; Qin, M.; Zhuang, X.; Li, X.; and Chen, H. 2024. InstructProtein: Aligning Human and Protein Language via Knowledge Instruction. Bangkok, Thailand: Association for Computational Linguistics.