Investigating the Power of Large Language Models for Advancements in Biomedical Research — From Quantitative Omics to Inductive and Deductive Reasoning

The way we get information has changed dramatically this year, from Google to using a large language model (LLM) like ChatGPT, which is now a built-in system-wide tool in the Windows 11, funded by Microsoft whose stock has increased 30 percent this year. Yet this change has not yet occurred significantly in the field of biomedical research [1], where PubMed/database inquiry still plays a major role in how people get information. I would like to have a reasoning system that people can use for medical research beyond current prevalent methods.

RNA fields have always been magic fields in the world of genomics: novel RNA vaccines have been developed to combat pandemics, and the discovery of glyco-RNA in cell membranes has left science in a state of shock[2]. It is noteworthy that the FDA has approved five siRNA drugs that specifically target mRNA for disease treatment [3]. The RNA molecular family encompasses various types, including mRNA, miRNA, lncRNA, and siRNA, which have all garnered significant attention and intensive study.

In the past 5 years, there has been a growing trend towards quantitative approaches to studying RNA. This has led to many omics sequencing projects, as well as a number of complementary computational or statistical methods for analysis. However, many of these methods are qualitative based on expression data. When you open a typical RNA database, you will see that the data is presented in a quantitative expression format: https://mirgenedb.org/browse/hsa)

Biomedical research should have not just quantitative omics methods, a reasoning language system might also be needed that can connect pieces of finding from studies. **I hypothesize that the LLM offers a promising solution that researchers can simply input their experimental findings into the model and gain valuable insights**. The model can list the reasons that may cause the observed phenomenon, identify downstream pathways that may be affected, and suggest the next steps to be taken. This approach can lead to the generation of new scientific hypotheses.

Like screenwriters now using LLM to write Netflix script with intended plots, the philosophy is with prompting instruction, LLM can query from network to output relevant results. I think this is extremely useful for biomedical research which has many confounding factors, so much data, like the topic RNA molecules. It is hard for a human brain to think based on a lot of factors.

I will collect papers that have studied each RNA molecule name, using prompt engineering to summarize the findings directly related to the RNA molecule to its biological meaning, which could be molecular, cellular, biochemical, or disease perspectives. I will connect them all together for a particular RNA molecule. The information of each RNA molecule will be chunked and transformed to store in a vectorbase with a separate index, and a fine-tuning question-answer based model would be generated. I would take advantage of new tools from the open-source community. The user can search information about RNA molecules and their biological meaning, and to access the latest research on the topic.

I will run multiple trials to test the parameters and performance. Literature chunking will be assigned and scored based on multiple categories. The score will be useful when creating the output, as higher vector scores should result in better retrieval-augmented generation (RAG). My experiments would include data source securing, chunking and indexing optimization, retrieval and generation prompt engineering. I will work with RNA researchers to test the performance.

Beside literature sources, different kinds of omics data will be separately trained. Here I propose to combine the LLM method with high dimension omics data. The omics expression data will have an assigned weighting factor (the importance of this molecule) before transformed and assigned. I hypothesize that this will have an unsurprising performance for retrieval-augmented generation (RAG) performance. A very useful scene would be when a user defines a query: What does it mean when X mRNA is silent, but Z is upregulated? The system can do a search based on information we stored on X, Z, or the relevant genes and have a prediction about the biological meaning.

My research also includes keeping up with the latest AI advances. Every day, new methods and tools are released in various areas of AI, such as running models with larger query tokens, running models on general PC, and training models on personal CPU, multi-modality applications. When I am faced with a question, I first look for an existing solution or a similar solution from another field that I can borrow. This is my philosophy for solving problems, and it saves time. If there is no solution that can be used or borrowed, then I will consider developing a first-of-its-kind solution. I pay attention to broad range of topics, from computer science to biology, from bioinformatics to AI, from visualization to medical imaging. **I will try to combine any interesting tools from any field to test if they can be used in my application**s. Taking an interdisciplinary approach can make all the difference. As my research is more like an engineering project, I am more focused on making the project useful, innovative, and beneficial to people, rather than developing elaborate theories or algorithms.

Intellectual Merit

This tool will be the first of its kind in the exciting RNA research field and a pioneer RAG application for medical research. This model will serve researchers, enabling them to generate new hypotheses and make inferences based on their findings. It will facilitate the connection of diverse knowledge, providing insights that cannot be obtained through search engine or online database. My model can also be used by biotech companies to develop more focused applications for their study molecule. **Alternatively, the more vital argument is to prove the language reasoning systems have the potential to avoid biases associated with the current omics based genetic-centric approaches[1]**. **As a result, researchers can access a wealth of information about RNA and overcome limitations in human-driven information retrieval.**

Broader Impacts

RNA molecules are a group that has an international society dedicated to its study, signaling its importance. I would like to present my projects at the annual meetings and find potential collaborators. Conference helps me to expand my knowledge and find new ideas. I would make a business card to follow up with attendees and request to try my products and provide feedback for improvement. I would ensure that the model remains auto-regular updated with the latest dataset. It is crucial for the success of our project. Additionally, I plan to attend AI conferences on CS and medicine field, find new translatable tools that can be utilized, and meet researchers and open-source community leaders whom we can collaborate with.

[1]"Prof. Nikolai Slavov on X: 'Currently, biomedical language is rather gene centric. This trend seems to have started in the 1960s, and perhaps it can be traced in part to Jim Watson and his self-promotion efforts.' [Twitter post]." Twitter, 21 June 2023, 12:55 PM, twitter.com/slavov_n/status/1671502123354271745.

[2] Some RNA molecules have unexpected sugar coating. HHMI. https://www.hhmi.org/news/some-rna-molecules-have-unexpected-sugar-coating

[3]Ahn IS, Kang C, Han J. Where should siRNAs go: applicable organs for siRNA drugs. Experimental & Molecular Medicine. 2023;55(7):1283-1292. doi:10.1038/s12276-023-00998-y