Please describe why this is the right tutorial to present at this conference and for the ISMB/ECCB community? Note: this is not the abstract.*

In the era of artificial intelligence, Large Language Models (LLMs) like ChatGPT are playing a crucial role in transforming biomedical data science research. Understanding and mastering these models has become indispensable for professionals operating in the interdisciplinary realm of biomedicine and informatics, such as the ISMB/ECCB community members. Our tutorial is designed to comprehensively cater to this pressing need of the community. Our tutorial provides a timely introduction and hands-on exercise in using LLMs to streamline biomedical data science research, including topics on academic literature exploration and bioinformatics programming. Furthermore, the tutorial's relevance extends to its contribution to enhancing the productivity of the ISMB/ECCB community members by enabling them to optimize their data handling and interpretation tasks, thus boosting their research productivity.

A Practical Introduction to Large Language Models in Biomedical Data Science Research

Proposal for a full-day virtual tutorial at ISMB 2024

Organizers/Speakers:
Robert Xiangru Tang, Yale University, USA.
Qiao Jin, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA.
Hufeng Zhou, Biostatstics Department, Harvard T. H. Chan School of Public Health, Harvard University, USA.
Shubo Tian, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA.
Zhiyong Lu, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA.
Mark Gerstein, Yale University, USA.

Abstract

Large Language Models (LLMs) like ChatGPT have exhibited remarkable capabilities in understanding and generating language across diverse disciplines. In the realm of biomedical data science, LLMs can significantly aid the processes of information accessibility, data analysis, and knowledge discovery. In this tutorial, we offer an introductory level hands-on guide to understanding and utilizing these LLMs in the field of biomedical data science. Our tutorial begins with leveling the learning ground by providing introductions to LLMs and Biomedical Data Science. Subsequently, we delve into the core applications of LLMs in biomedical data science via retrieval-augmented generation, database functionalities, and code generation. To facilitate thought-provoking discussions, pertinent case studies will be discussed, emphasizing how to harness the power of LLMs to bridge the gap between technical feasibility and practical utility in biomedical data science. Furthermore, hands-on exercises are included to enable participants to apply their learning in real-time. Participants will also get acquainted with OpenAI's ChatGPT and open-source LLMs, as well as their design, use cases, limitations, and prospects.

Our topics include:

- Understanding Large Language Models (LLMs) and their evolution from RNNs, LSTM to Transformers and GPT family.
- Introducing Biomedical Data Science and the challenges faced in the field.
- In-depth interaction with OpenAI’s ChatGPT, learning about its overview, capabilities, and implementation, focusing on Chain-of-Thought Prompting, Retrieval-Augmented Generation, Text2SQL, and Language Agents.
- Analyze detailed real-world case studies: Retrieval over PubMed and Code Generation.
- Guided hands-on exercises using provided datasets and problem statements for practical understanding and implementation.

This tutorial is an activity of the proposed ISCB COSI on Text Mining and the Impact of Genomic Variation on Function (IGVF) Consortium.

Learning Objectives

- Familiarizing with the key aspects of large-scale biomedical data.
- Leveraging LLMs to handle and interpret vast amounts of biomedical data.
- Learning cutting-edge research topics from two invited talks.
- Utilizing OpenAI APIs for GPTs and open-source LLMs in Python.
- Integrating LLMs to enhance their coding efficiency in bioinformatics.
- Deploying LLMs for biomedical question-answering and academic literature exploration.

By the end of this tutorial, attendees will be equipped with the intro-level knowledge and practical skills to harness the capabilities of LLMs in their research.

Short Promotional Blurb for Promotion

Unleash the power of Large Language Models (LLMs) to transform your Biomedical Data Science Research. Equip yourself with the skills of using GPT-3.5, GPT-4, and more to delve deep into the sea of Biomedical Data. Learn to code, and perform QA and academic literature exploration with tips, tricks, and insights from the leaders in the field!

Maximum Number of Attendees Participating

40

Draft Schedule (Full Day):

Slot	Theme	Speaker
09:00 - 10:30 Session 1 Introduction & Basics		
09:00 - 09:10	Overview and Welcome	Robert Tang
09:10 - 09:40	Introduction to LLMs with a focus on Biomedical Data Science	Shubo Tian
09:40 - 09:10	How to use GPT-3.5 and GPT-4 with Python	Qiao Jin
10:10 - 10:30	How to use Open-source	Robert Tang

	LLMs with Python	
10:30 - 10:45 Break		
10:45 - 12:00 Session 2 Application of LLMs in Biomedical Data Science		
10:45 - 11:10	Database Query with LLMs	Hufeng Zhou
11:10 - 11:35	Retrieval-augmented Generation with Large Language Models	Qiao Jin
11:35 -12:00	Code generation in Bioinformatics	Robert Tang
12:00 - 13:00 Lunch		
13:00 - 14:30 Session 3: Advanced Topics		
13:00 - 13:45	Large Language Models for Biomedicine: from PubMed Search to Clinical Trial Matching	Zhiyong Lu
13:45 - 14:30	AI in Biomedicine: Developing Representations of Disease-Relevant Molecules	Mark Gerstein
14:30 - 14:45 Break		
14:45 - 15:55 Session 4: Hands-on Exercise		
14:45 - 15:10	Integrating Biomedical Data Database Development with LLMs	Hufeng Zhou
15:10 - 15:35	Querying PubMed with RAG to answer biomedical questions with GPT-4	Qiao Jin
15:35 - 15:55	Code generation in Bioinformatics with Open-source LLMs	Robert Tang
15:55 - 16:00 Closing Remarks		

Hands-on Content

In our tutorial, we offer three sections of hands-on content, incorporated via practical exercises and real-world case studies that encourage participants to apply the theoretical knowledge they gain.

- In-depth interaction with OpenAI’s ChatGPT and open-source LLMs: This part of the tutorial engages participants in practical interaction with ChatGPT to understand its functionalities and capabilities. Participants will learn about Chain-of-Thought Prompting, Retrieval-Augmented Generation, Text2SQL, and Code Generation through demonstration and practice.
- Real-world case studies: We will analyze detailed real-world case studies, emphasizing Retrieval over PubMed and Code Generation. These case studies have been selected to inspire participants regarding ways LLMs can be effectively used to enhance biomedical research. After each case study discussion, participants will be encouraged to reflect on the application of concepts in their research scenarios.
- Guided hands-on exercises: Throughout the tutorial, participants will engage with multiple hands-on exercises. These exercises will leverage provided datasets and problem statements that simulate real-world biomedical research challenges. Participants will be expected to apply the concepts learned about LLMs through Python to address these problems practically and efficiently. These guided hands-on exercises will solidify participants' understanding of LLMs and their application in the biomedical data science field.

Intended audience and level

This tutorial is designed for graduate students, researchers, data analysts, and practitioners in the domains of bioinformatics, computational biology, and biomedical informatics who are seeking to harness the potential of Large Language Models (LLMs) in their work. The didactic content would be chiefly beneficial for individuals who are keen on enhancing the breadth and depth of their analytical skills.

While the focus of the workshop lies in catering to beginners or users with little experience in LLMs, intermediates will find the advanced topics and in-depth case studies enriching as well. Participants should ideally possess a basic understanding of Python programming and machine learning concepts. Preliminary experience with Linux-based operating systems or interacting with APIs would provide an added advantage but is not a prerequisite.

Our discussion on using OpenAI's ChatGPT and other open-source LLMs, along with hands-on exercises and case studies, will offer an immersive learning experience that spans theory and practice. Researchers looking to streamline their data analysis processes and improve the efficiency and accuracy of their results will find this tutorial particularly useful.

Relevant resources and tutorial materials for hands-on activities will be shared online before the commencement of the tutorial, ensuring an unhampered learning experience for all attendees.