MSc thesis project: Large language models to augment literature surveys in engineering acoustics

Background: Remarkable advances in large language models (LLMs) and generative pre-trained transformer (GPT) methods[1] offer a unique opportunity to aid humans in performing literature surveys of hundreds (and potentially thousands) of papers. As shown in Fig. 1, a Scopus search on data-driven methods in acoustics demonstrates that the number of articles in the first half of 2023 exceeded the total number of articles in 2019. Given this growth in the literature, the acoustics community faces the limitations of traditional survey methods.

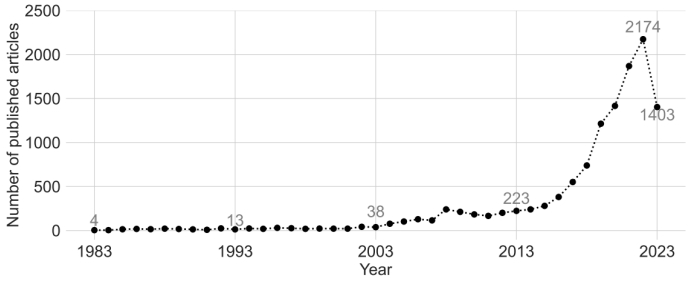

Figure 1. Four decades of articles on applications of data-driven methods in acoustics. Scopus search on August 2, 2023.

Studies on acoustics corpora with few (and zero-) shot approaches have been done. We have tested GPT-3.5 with a corpus of 116 articles on data-driven speech-enhancement methods[2]. More recently, we have looked at the capabilities of Databricks' Dolly and HugginFace's Langchain models for batch-processing approach, querying 500 abstracts at a time using OpenAI agents. There is great potential for improvements at the prompting level, post-processing, and further analysis of the responses.

Description: The goal of the MSc project is to design, implement, and assess the capabilities of LLMs (e.g., GPT-3.5/4.0, Dolly, Langchain, Huggingface agents) for surveying a corpus of hundreds of articles in data-driven methods in acoustics. The project is a collaboration between the Marcus Wallenberg Laboratory for Sound and Vibration Research (MWL), Dept. Engineering Mechanics, and Department of Intelligent Systems, Division of Robotics, Perception, and Learning. To start, the student will perform a literature survey of current approaches in automatic summarization and multi-document query. After that, the student will build upon existing LLMs, deploy them, and analyze the results. The performance evaluation of the models will be carried out through commonly used natural language processing metrics and compared with expert auscultations. The student will count on a supervisor from EECS and domain expertise on acoustics from the supervisors at SCI.

Student qualifications: The student should preferably have a background in computer science, data science, machine learning, or the like. Experience with Python programming, repositories, data lakes, and similar cloud-based services (e.g., Microsoft Azure, databricks) is valuable. Experience with natural language processing and large language models is a big plus.

Starting date/Project length: VT24 / ASAP. Up to 6 months.

Supervisors: Elias Zea (zea@kth.se), Shiva Sander Tavallaey (shiva.sander-tavallaey@se.abb.com),

[1] C. Stokel-Walker, R.V. Noorden (2023) "*What ChatGPT and generative AI mean for science*," Nature **614**, 214–216.

[22] A. dos Santos, J. Pereira, R. Nogeuira, B. Masiero, S. Sander-Tavallaey, E. Zea (2023), "*An experiment on an automated literature survey of data-driven speech enhancement methods*," arXiv:2310.06260, https://arxiv.org/abs/2310.06260.