Discovery of Novel STEM Documents

1st Tofara Moyo
Bulawayo , Zimbabwe
tofaramoyo@gmail.com, Mazusa AI

***Abstract*—We present a novel scientific document discovery system inspired by molecular chemistry and AI-driven drug discovery. Our approach treats document tokens as atomic units, which are combined to form "molecular" representations of mathematical documents. We employ a probabilistic framework to maximize the likelihood of forming coherent mathematical documents while minimizing the probability of random token combinations and non-STEM document tokens.To achieve this, we develop a token embedding scheme that maps property vectors to a musical keyboard, effectively representing each token as a musical chord. We further differentiate between STEM and non-STEM documents by introducing a harmonic constraint on adjacent nodes in document graphs. Specifically, STEM documents are characterized by polyphonic harmonization of adjacent node vectors, whereas non-STEM documents exhibit dissonant relationships.Our system integrates a graph neural network/transformer decoder architecture, trained end-to-end to generate STEM documents from input graphs. This innovative approach has the potential to revolutionize scientific document discovery and retrieval.**

I. Introduction

Recent advancements in large language models have yielded impressive results in natural language processing tasks. Transformer-based neural networks trained on vast corpora of text, such as the internet, have demonstrated remarkable accuracy in answering questions across diverse topics. However, a notable knowledge gap persists in the realms of science and mathematics.Despite their capabilities, current language models struggle to engage with scientific and mathematical content in a meaningful way. While they can converse about scientific information using varied linguistic permutations, they are unable to generate novel scientific articles or contribute original ideas. Furthermore, these models fail to perform even basic mathematical calculations, let alone conceive innovative mathematical concepts.This limitation underscores the need for specialized language models that can be trained on scientific corpora to produce original research articles and advance the frontiers of scientific knowledge.

This paper proposes a novel approach for developing a language model capable of generating original scientific content de novo.

II. STEM producer

In developing our system for generating scientific content, we drew inspiration from the fundamental principles governing molecular formation in chemistry. Atoms, as the basic building blocks of matter, possess inherent local properties, such as atomic number and quantum physical attributes, which dictate their propensity to form connections with other atoms and ultimately give rise to complex molecules.

This atomic-molecular paradigm bears intriguing parallels with the axiomatic foundations of mathematical proof systems. Mathematical proofs are constructed from a set of axioms and governed by rules of inference, which collectively enable the derivation of theorems and propositions.

Our system leverages this conceptual analogy by representing STEM documents as graphs comprising words and alphanumeric characters, herein referred to as tokens. Each token is associated with a unique property vector, which encapsulates its local properties and influences the likelihood and characteristics of the graph it forms. This innovative approach enables the capture of intricate relationships between tokens and the generation of coherent scientific content.

Our approach draws inspiration from the work of Moyo et al, who proposed a novel method for regularizing spiking neural networks using music theory. In their framework, neurons in each layer were mapped to a musical keyboard, and at each time step, the firing nodes were arranged to harmonize, mimicking the keys they represented.

We adapt this concept to our architecture. Our model consists of a Graph neural Variational Autoendoer that conditions the latent encoding on text input. The encoded text is then used to generate a graph, which is fed into a traditional graph neural network (GNN) to produce an encoding. This encoding is subsequently utilized by a transformer decoder to generate the final text output.

In our implementation, we employ a dissonance calculation plugin, the "Dissonant" plugin , to evaluate the perceptual dissonance of node combinations in the graph an within nodes. This dissonance metric informs the harmonization of nodes, ensuring that the generated graph adheres to a harmonious structure.

To effectively represent the subset of text documents pertinent to STEM topics, we propose leveraging a specialized dataset comprising STEM-focused Wikipedia articles, arXiv papers, and mathematical proofs sourced from repositories such as ProofWiki. In conjunction with this STEM-centric dataset, we also employ a traditional text corpus to provide a broader linguistic context. However, we adopt differing training objectives for these two datasets, acknowledging the distinct characteristics and content of each. This dual-dataset approach enables our model to differentiate between both STEM-specific language and the general patterns of natural language.

As mentioned , each token will be associated with a property vector. We will not define the properties, but instead map the indices of the vector to a musical keyboard. Then we train the system in such a way that each token gets associated with a different musical chord. Then only in those documents based off STEM topics do we arrange for adjacent nodes vectors chords to harmonize. So we train for harmonization within each tokens property vector first. Then with a graph in mind , we train for the graphs of STEM documents to have adjacent nodes that have property vectors that harmonize while the regular documents are encouraged to have inharmonious adjacent nodes.This hierarchical training approach enables our model to capture both the intrinsic properties of individual tokens and the structural relationships governing STEM documents.

III. Methodology

Our proposed system comprises three synergistic components, designed to operate in tandem to generate coherent STEM text. The architecture consists of a Graph Neural Variational Autoencoder (GN-VAE), responsible for learning a probabilistic representation of the input graph structure.A traditional Graph Neural Network (GNN), which processes the output of the GN-VAE to refine the graph representation and capture higher-order structural relationships. Transformer Decoder, which takes the output of the GNN and generates the final text output, leveraging the structural and semantic information encoded in the graph representation.The integration of these three components enables our system to effectively capture the complex relationships and structures inherent in STEM text, and generate coherent and informative text outputs.

We propose training the integrated Graph Neural Network (GNN) and Transformer Decoder architecture in an end-to-end manner. This training paradigm enables the system to jointly optimize the representation learning and text generation processes.Consequently, the model will automatically assign a unique musical chord representation to each token within our predefined token set. This harmonization-based token embedding scheme allows the model to capture intricate structural relationships between tokens and generate coherent text outputs.

Our training protocol involves the following steps:

Graph Initialization: For each document in our training corpus, we generate a random graph comprising tokens present in the document.

Loss Function Formulation: We define a loss function comprising two primary components: (a) Mean Squared Error (MSE) between predicted and actual token representations, and (b) a dissonance term, which penalizes inharmonious token combinations.It is also designed to treat the property vectors of the tokens as weights to be updated through the learning process.

Document-Specific Loss Function Modification: When training on STEM documents, we augment the loss function with an additional term that penalizes dissonance between connected nodes in the graph. Conversely, when training on regular documents, we modify this term to reward consonance between connected nodes.

Graph Pruning: Following each training epoch, we revisit the graphs and prune connections between nodes in STEM document graphs that exhibit low consonance. Conversely, we prune connections between nodes in regular document graphs that exhibit high consonance. This training protocol enables our model to learn harmonious relationships between tokens in STEM documents and inharmonious relationships in regular documents.

To circumvent the computational infeasibility of utilizing fully connected graphs, we employ an iterative graph randomization strategy. Specifically, in each subsequent training epoch, we randomly reinitialize the connections between nodes in the graph. We then repeat the training protocol outlined in the previous section, using the newly randomized graph structure.This iterative approach enables us to efficiently explore the vast space of possible graph connections, while avoiding the prohibitive computational costs associated with fully connected graphs.

Upon completing the initial training phase, we proceed to train the Graph Variational Autoencoder (GVAE) using the STEM graphs generated using the probabilities of the nodes as connections.For each STEM document we create a fully connected graph. Then after computing the harmony of adjacent nodes we keep those connections with high consonance and prune those with low. These graphs possess property vectors associated with each node, which encapsulate the musical chord representations.The next step would be to retrain the graph neural network with these pruned graphs.Our objective in training the GVAE is to enable the model to discover the underlying structural principle of ordering nodes based on

consonance, as dictated by the music group. By learning this simple yet effective rule, the GVAE is expected to generate more coherent and scientifically valid documents, particularly in the realms of mathematics and science.

IV. Conclusions

In this paper, we presented a novel scientific document discovery system, inspired by the principles of molecular chemistry and AI-driven drug discovery. Our approach leverages a novel token embedding scheme, mapping property vectors to a musical keyboard, and introduces a harmonic constraint to differentiate between STEM and non-STEM documents.

The proposed system integrates a graph neural network/transformer decoder architecture, trained end-to-end to generate STEM documents from input graphs derived from the graph variational autoencoder's output. This innovative approach has the potential to revolutionize scientific document discovery and retrieval.

Our work demonstrates the power of interdisciplinary research, combining concepts from chemistry, music theory, and artificial intelligence to tackle complex challenges in scientific document analysis. Future research directions include exploring the application of this framework to other domains, such as patent analysis and biomedical literature mining.

Ultimately, our system has the potential to facilitate the discovery of novel scientific concepts, accelerate the pace of research, and drive innovation in various fields of study.

References

[1] Comparative Analysis of CHATGPT and the evolution of language models Oluwatosin Ogundare, Gustavo Quiros Araya

[2] Mathematics and group theory in music Athanase Papadopoulos (IRMA)

[3] A Group-theoretic view of Music-Dimitris ChatzisDimitris Chatzis

[4] Heaps' Law and Vocabulary Richness in the History of Classical Music Harmony-Authors: Serra-Peralta, Marc, Serrà, Joan, Corral, Álvaro

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.

[6] Structuring Concept Space with the Musical Circle of Fifths by Utilizing Music Grammar Based Activations Tofara Moyo https://arxiv.org/abs/2403.00790