Leveraging Grounded Large Language Models to Automate Educational Presentation Generation

Eric Xie, Guangzhi Xiong, Haolin Yang, Aidong Zhang
University of Virginia
Charlottesville, VA 22903
{jrg4wx, hhu4zu, upc6ps, aidong}@virginia.edu

Abstract

Large Language Models (LLMs) have shown great potential in education, which may significantly facilitate course preparation from making quiz questions to automatically evaluating student answers. By helping educators quickly generate high-quality educational content, LLMs enable an increased focus on student engagement, lesson planning, and personalized instruction, ultimately enhancing the overall learning experience. While slide preparation is a crucial step in education, which helps instructors present the course in an organized way, there have been few attempts at using LLMs for slide generation. Due to the hallucination problem of LLMs and the requirement of accurate knowledge in education, there is a distinct lack of LLM tools that generate presentations tailored for education, especially in specific domains such as biomedicine. To address this gap, we design a new framework to accelerate and automate the slide preparation step in biomedical education using knowledge-enhanced LLMs. Specifically, we leverage the code generation capabilities of LLMs to bridge the gap between modalities of texts and slides in presentation. The retrieval-augmented generation (RAG) is also incorporated into our framework to enhance the slide generation with external knowledge bases and ground the generated content with traceable sources. Our experiments demonstrate the utility of our framework in terms of relevance and depth, which reflect the potential of LLMs in facilitating slide preparation for education.

1 Introduction

The rapid development of artificial intelligence (AI) technologies has provided great opportunities and changes to various areas such as finance, medicine, and education Bahoo et al. [2024], Kitsios et al. [2023], Fitria [2021], Holmes et al. [2019]. One of the major breakthroughs in AI development is the introduction of large language models (LLMs), which show great capabilities in a variety of tasks in different domains Achiam et al. [2023], Touvron et al. [2023], Team et al. [2023]. With the ability to follow user instructions Ouyang et al. [2022] and learn from the context Brown [2020], LLMs have demonstrated their potential in facilitating educators in different levels of teaching Elkins et al. [2024], Agrawal et al. [2024], Fagbohun et al. [2024].

Slide presentation plays a crucial role in education, as it helps instructors present information in a clear, structured, and engaging way Alley and Neeley [2005], Mayer [2002], Bartsch and Cobern [2003]. However, creating high-quality slides can be a challenging and time-consuming task, requiring educators to distill large volumes of content into concise, visually appealing formats. By leveraging their advanced capabilities in text generation and summarization, LLMs present a significant opportunity to automate and streamline the slide preparation process. With instructional objectives from the educators, LLMs can automatically generate content that is well-organized, concise, and tailored to specific educational needs.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

While LLMs offer significant potential for automating slide preparation, they carry the risk of generating inaccurate or misleading content, which can be fatal in education Bender et al. [2021], Bommasani et al. [2021]. LLMs may produce information that appears credible but lacks factual accuracy, especially in specialized fields such as biomedicine, where the models may not possess enough domain-specific knowledge to handle the complexity of the subject during slide generation Ahmad et al. [2023], Arighi et al. [2023]. Moreover, biomedicine is a rapidly evolving field, with new research and developments emerging frequently, making it challenging for LLMs to stay updated with the latest knowledge Collins et al. [2021], Flier [2023], Cremin et al. [2022].

To address the problems mentioned above and leverage the capabilities of LLMs to facilitate slide preparation in education, we propose a retrieval-augmented slide generation framework, which explicitly extracts relevant information for given instructions from external knowledge bases and grounds the generated slides in authoritative and reliable domain-specific knowledge. By using the Beamer[1] package in LaTeX[2] as the medium to bridge text input and visual output, we leverage the code generation capability of LLMs to automate the slide preparation process by generating the LaTeX scripts. Our framework also allows collaboration between instructors and LLMs, facilitating personalized slide preparation by giving detailed instructions to the models. We evaluate our framework on 9 different biomedical topics with various complexities including introductory, intermediate, and advanced ones. To assess the generated slides provided by GPT-4o and GPT-4o-mini, 4 human annotators are incorporated to examine the slides from the perspective of relevance, depth, and overall assessment. Our results demonstrate the effectiveness of LLMs in generating educational slides, with high relevance and depth in various settings. Moreover, we show that the incorporation of retrieval-augmented generation (RAG) in our proposed framework significantly improves the depth of generated slides by adding more details to the content. We also illustrate how our system can collaborate with human instructors to change the final output flexibly. Our findings reveal the huge potential of LLMs in slide generation for educational purposes, which could facilitate course preparation of instructors and help provide high-quality materials to students.

2 Related Work

2.1 Slide Generation and Editing

The task of automating the generation of presentation slides has been an area of growing interest, particularly for scientific and technical papers. Early work in this area focused on extractive methods. Sefid et al. [2019] proposed a method based on the SummaRuNNer model, adapting it for scientific papers. Their approach uses a windowed labeling ranking system, combining semantic and lexical features within a sentence window to measure the importance and novelty of sentences.

Other researchers have explored various techniques for slide generation. Hu and Wan [2015] developed PPSGen, a framework that uses Support Vector Regressors and Integer Linear Programming (ILP) to rank and select important sentences. Wang et al. [2017] took a different approach, focusing on extracting phrases from papers and learning hierarchical relationships between them to structure bullet points. More recent work has begun to leverage deep learning techniques. Sefid et al. [2021] extended their previous work by incorporating a more comprehensive list of surface features, considering the semantic meaning of sentences, and using contextual information for ranking. Their method combines feature-based and deep neural network approaches for sentence scoring, followed by ILP for summary construction.

The challenge of working with longer documents has also been addressed. Gupta [2023] explored the use of large language models with extended token limits, such as Longformer-Encoder-Decoder and BIGBIRD-Pegasus, to handle the full length of scientific papers. This approach yielded promising results, particularly when training on section-slide pairs, showing improved coherence as measured by R2 and RL scores.

[1] `https://ctan.org/pkg/beamer`

[2] `https://www.latex-project.org`

2.2 Retrieval-augmented Generation

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [2020], aims to enhance the performance of language models on tasks requiring extensive knowledge by incorporating relevant retrieved information. This approach offers two significant advantages: it reduces the likelihood of AI-generated falsehoods by grounding the model's outputs in specific contexts, and it allows for the inclusion of current information that may not be part of the model's original training data. Since its inception, numerous researchers have built upon and refined the original RAG concept Borgeaud et al. [2022], Ram et al. [2023], Gao et al. [2023], Jiang et al. [2023], Mialon et al. [2023].

In the biomedical domain, several studies have explored the potential of large language models (LLMs) enhanced with RAG to improve literature searches and support clinical decision-making processes Frisoni et al. [2022], Naik et al. [2022], Jin et al. [2023], Lála et al. [2023], Zakka et al. [2024], Jeong et al. [2024], Wang et al. [2023], Xiong et al. [2024]. However, the potential of RAG on in biomedical education is still under-explored. In this study, we leverage the advantages of RAG to ground the slide generation of LLMs in well-documented scientific knowledge.

2.3 Biomedical Education

The integration of artificial intelligence (AI) in biomedical education is transforming how instructional content is generated and delivered. In recent years, AI has been utilized to create adaptive learning tools, assessments, and personalized content for students, particularly in medical fields where rapid advancements in knowledge require innovative educational approaches Kasneci et al. [2023], Elkins et al. [2024], Mir et al. [2023].

Sridharan and Sequeira [2024] conducted a proof-of-concept study exploring the application of generative AI tools in pharmacology education. They demonstrated the capability of AI in generating specific learning outcomes (SLOs), various types of test items, and test standard-setting parameters. Mir et al. [2023] provide a broader perspective on AI's role in medical education. They identify several key applications, including Virtual Inquiry Systems, Medical Distance Learning and Management, and recording teaching videos. Their work underscores AI's potential to address various educational challenges, from language processing to cognitive modeling. Veras et al. [2023] are conducting a randomized controlled trial to investigate the usability and efficacy of AI chatbots, specifically ChatGPT, as a supplementary learning tool for health sciences students.

While these studies demonstrate the growing integration of AI in various aspects of biomedical education, there remains a notable gap in the literature regarding the application of AI for generating teaching slides in the biomedical field.

2.4 Artificial Intelligence in Education

Artificial Intelligence has been increasingly integrated into educational contexts, revolutionizing teaching and learning processes. LLMs have shown particular promise in this domain Alsafari et al. [2024], Moore et al. [2023], Kasneci et al. [2023]. One notable area is Question Generation (QG), where AI is used to generate educational quizzes and questions. For example, Elkins et al. [2024] demonstrate that LLM-based question generation can produce quizzes as effective as those written by teachers. In fact, the automatically generated questions were shown to be of equal or higher quality, reducing the time teachers spend on creating assessments while maintaining educational integrity. Similarly, Agrawal et al. [2024] developed CyberQ, a system that uses knowledge graph-augmented LLMs to generate questions and answers for cybersecurity education. This approach demonstrates how AI can create tailored educational content in specialized fields. Interactive learning systems powered by AI have also gained traction. Chen et al. [2021] created a chatbot-based question-answering system for students, showcasing AI's potential in providing personalized learning experiences. Similarly, Dan et al. [2023] introduced EduChat, a large-scale LLM-based chatbot system for intelligent education in Chinese middle and high school curricula. While LLMs hold great potential to positively transform educational practices and ultimately student educational outcomes, it is important to remember that their continued integration in the field of education should be approached with a balanced perspective that considers both their benefits and the inherent limitations Huber et al. [2024], Kasneci et al. [2023], Stamper et al. [2024].

3 Methodology

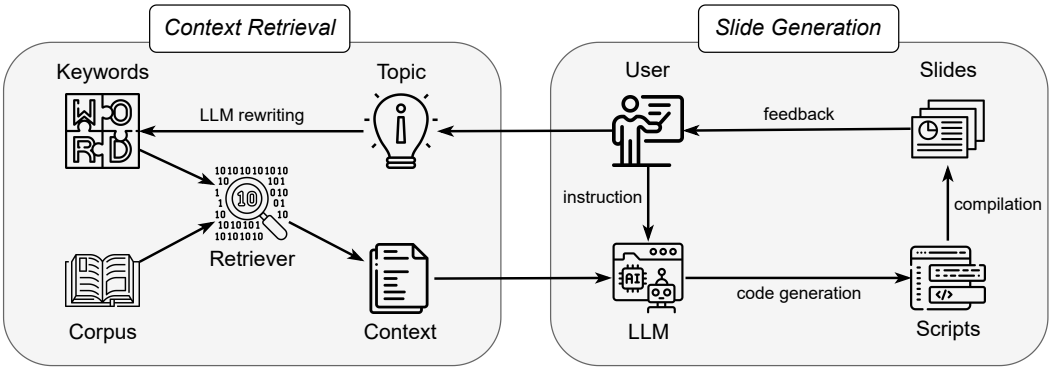

Figure 1: An overview of the complete slide generation pipeline, highlighting the Context Retrieval and Slide Generation components and their respective constituents.

In this section, we introduce the two major components of our slide generation framework: context retrieval and slide generation, depicted in Figure 1. We additionally explore the option of collaborative development between the instructor and the system following the generation of a presentation. The context retrieval component automatically compiles sections relevant to the given topic, extracted from within a collection of biomedical textbooks. This provides valuable context that grounds the output of the generation model in the existing biomedical knowledge. The slide generation component then takes the retrieved information and transforms it into structured presentation slides, highlighting the key information in a clear and concise manner.

3.1 Context Retrieval

Retrieval-augmented generation Lewis et al. [2020] (RAG) is crucial for grounding the outputs of the slide generation model by leveraging information from domain-specific sources, ensuring that the content is accurate and contextually relevant. Grounding the model's output in established sources helps mitigate the risk of hallucinations. This reduces the amount of incorrect or misleading information by ensuring content is tied to verifiable data Béchard and Ayala [2024]. Figure 2 displays an example of verifiable content within an LLM generated presentation compared against the textbook source material. As shown in the example, the model will identify, extract, then summarize relevant facts taken from within the provided context and add the result to the generated slides. Furthermore, the model will ensure these facts are referenced properly, creating a slide of references as dictated

by the generation prompt shown in Figure 3. The end result is a set of slides that conveys not only accurate, but sourceable information contextually grounded using the retrieval corpus.

We source our content from a collection of biomedical textbooks gathered by Jin et al. [2021], which provides crucial domain-specific information to the model. Since textbooks are a primary academic resource for most classrooms, this aligns our model with existing study materials and further demonstrates its practicality in real-world educational settings. To further enhance the relevance of the content, we select BM25 Robertson et al. [2009], a commonly used lexicon-based text retriever, to pinpoint and extract only the most pertinent sections across the textbooks within the retrieval corpus. Before entering the user instruction into the text retriever, we rewrite it by generating key terms for the instruction using LLMs to ensure the retrieval process targets the most relevant material. While our current focus is the biomedical field, this framework can easily be adapted to other domains due to the interchangeability of the corpus. The retrieved information can easily be adjusted to fit the personal needs of students or instructors by specifying the keywords to search. In addition to domain-specific textbooks, other information sources such as recent research papers could be substituted in to search for the latest knowledge without affecting the rest of the framework.

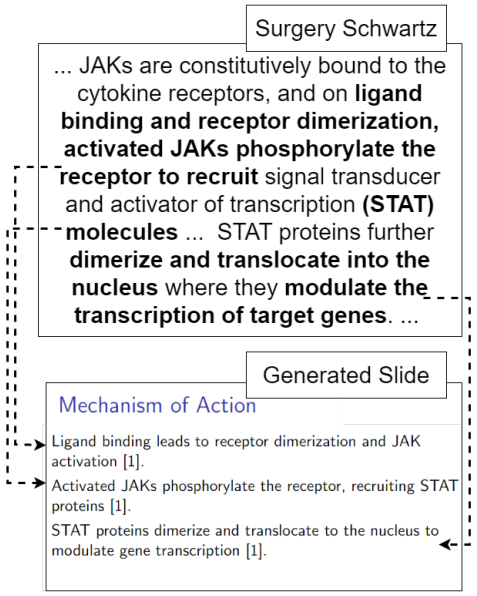

Figure 2: Example of an LLM-generated slide alongside a segment of the provided context between source material (*Schwartz's Principles of Surgery* for [1]). This showcases the pipeline's ability to extract and display relevant information from the given context.

Prompt Template for LLM Slide Generation Given Context

System Prompt: You are an expert at creating presentations for advanced courses tasked with generating a LaTeX script using the Beamer package to generate a presentation.

User Prompt: You are an expert tasked with generating a set of slides for a presentation for a graduate level course on the given topic based on the given content.

Make sure ALL information on the slides are supported using the given snippets from textbooks, citing the source beside every bullet point, but you do NOT need to use all snippets. Cite where each fact is obtained from on each slide: [list of textbook titles].

On the last slide of the presentation, ensure all uses of \\cite have proper references to the corresponding paper title using \\bibitem{{Paper Title}}, regardless of the section. The first slide after the title should roadmap the rest of the presentation, and you must ensure that ALL topics in your roadmap are covered in the presentation.

Figure 3: Prompt template used to generate presentation slides given retrieved context. "[list of textbook titles]" is automatically replaced with a list of the titles of the textbooks from which the snippets were acquired.

3.2 Slide Generation

This pipeline leverages the powerful code generation capabilities of modern LLMs Jiang et al. [2024] to automatically produce LaTeX scripts, enabling efficient and accurate creation of well-structured and customizable slides. As our slide generation medium, we select LaTeX, a typesetting system widely used for academic and scientific documents, along with Beamer, a LaTeX package specifically designed for creating presentation slides. This combination offers robust support for technical formatting and a flexible document structure, making it ideal for scientific and academic content. The flexibility of LaTeX and Beamer allows for superior handling of technical content, greater formatting control, and seamless integration with citation systems, making it especially useful for academic presentations when compared to traditional slide-making tools. Figure 3 displays the specific prompt template used for generating presentation slides given the retrieved context. This prompt is applicable

across various academic fields and provides a structured approach to ensure the generated slides align with the overall presentation goals.

Once generated, the created slides can easily be adjusted to accommodate specific content changes, design preferences, or formatting requirements. Since the output is in LaTeX, the instructor can collaborate with the model to further refine the presentation. After the initial generation, the instructor has the option to ask the model to make direct adjustments to the layout of the presentation, such as changing the theme, colors, and fonts, or suggesting additional information, descriptions, or potential images and figures to include. Alternatively, the instructor has the option to make their preferred modifications themselves, allowing for full control over the content and design of the presentation to suit the needs of the class. An example of this process is shown in Figure 4, illustrating how the slides can be transformed through collaboration between the instructor and the model.

4 Experiments

4.1 Experimental Settings

To evaluate the capability of our framework in generating effective biomedical presentations, we generate slides for 9 topics of varying complexity levels across different model configurations for human evaluation. Specifically, we use the GPT-4o and GPT-4o mini models, with and without contextual information. The selected topics within the biomedical field are categorized into three levels of complexity: introductory, intermediate, and advanced.

We evaluate the model on three introductory-level topics: “immune cells," “the nervous system," and “nucleic acids." Topics classified at the introductory level represent foundational knowledge that any graduate-level biomedical student is expected to understand, reflecting the content typically covered in introductory courses at the graduate level.

Next, the three intermediate level topics we evaluate the model on are “the mechanisms of antigen-presenting cells," “neurotransmitters," and “transcription regulation in eukaryotes." These intermediate topics delve into more specialized content, requiring a deeper understanding of specific processes and interactions within the biomedical field. These topics represent content typically covered in more advanced courses or research seminars, most applicable to students with a focus on that particular sub-field.

Lastly, we select “ clinical application of dendritic cells," “therapeutic effectiveness of branched-chain amino acids," and “transcriptional cofactors and post-translational modifications" as the three advanced-level topics. These topics are highly specialized and are expected to be encountered by students engaging in focused research or specialized academic work within the field. These topics require a comprehensive understanding of specific methodologies or processes, forming the foundation for conducting independent research or contributing to ongoing studies in the discipline. For example, a presentation on the clinical application of dendritic cells covers a topic that remains an active area of ongoing research.

The generated presentations are evaluated by a panel of four human reviewers who assess the content from the perspective of students in the biomedical field. Reviewers evaluated the presentations using two key criteria: relevance and depth. Relevance represents how useful each subtopic explained within a presentation is in contributing to the understanding of the overall topic. It assesses the significance of the subtopics within the presentation in explaining the main subject, irrespective of how thoroughly the subtopic is covered. Depth, on the other hand, measures how comprehensively each subtopic is explained, regardless of its relevance to the overall topic. This criterion evaluates the level of detail and the depth of understanding presented for each subtopic, focusing on the richness of the explanation provided. All metrics range from 1 to 5, representing the least to the greatest significance of the quality assessed.

There exist several alternative metrics to assess the quality and effectiveness of presentations, such as clarity and organization, engagement and visual appeal, or consistency in formatting. While the aesthetic qualities and organization of presentations can have a significant impact on slide quality, as displayed in Figure 4, we find that the collaborative development between the model and an instructor enabled by our framework can ensure that the overall structure and appearance of the presentation can be tailored to each students’ preferences. In this example, the model is instructed to provide and implement three different ways to emphasize the key information within the initial generated

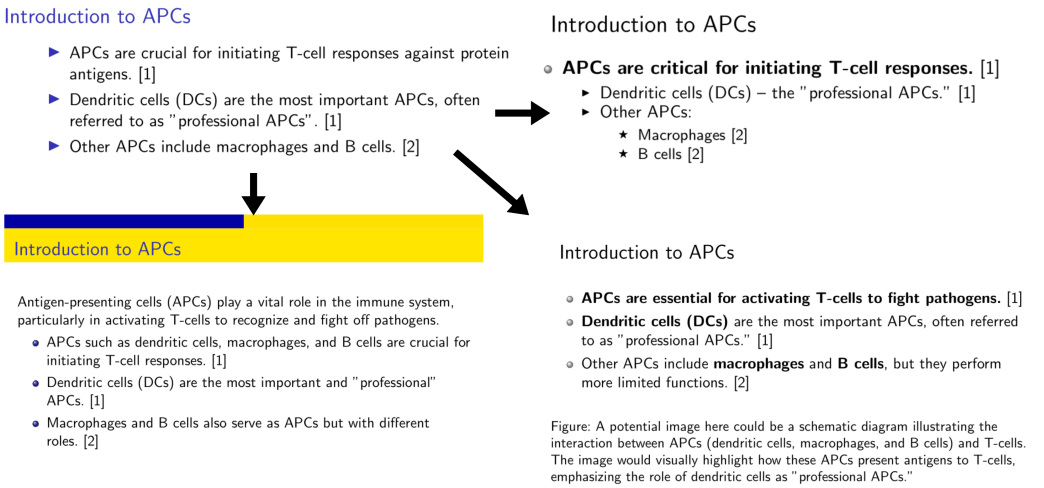

Figure 4: Example of a slide being enhanced through collaborative development, showcasing different ways an initial slide (top-left) can be improved through model-instructor interaction. These variations demonstrate improvements in the structure, emphasis, and advice on visual elements, highlighting how collaboration with the model can result in a presentation that offers a more effective and customized learning experience. [1] and [2] refer to *Schwartz's Principles of Surgery* and *Janeway's Immunobiology*, respectively.

slide (top-left) while simultaneously adding a theme or coloring scheme to the slide. In response, the model enhances the slide using three different strategies. The top-right slide changes the bullet structure, grouping related items and utilizing a sub-bullet structure, while the bottom-left slide adds a brief overview preceding the bullet points. The bottom-right slide enhances the presentation further by identifying then highlighting key information with bold text and incorporating a description of a supporting image to insert. Overall, the malleability of the slides generated using this framework allows for easy adjustments through collaborative development, causing more subjective metrics, such as organization or aesthetic quality, to be less critical to measure.

4.2 Main Results

Topic Complexity	Metric	GPT-4o Mini		GPT-4o	
		No Context	Context	No Context	Context
Introductory	**Relevance**	4.833 ±0.327	4.917 ±0.289	4.958 ±0.144	4.875 ±0.311
	Depth	3.792 ±0.450	4.542 ±0.396	3.708 ±0.656	4.208 ±0.582
	Overall	4.104 ±0.361	**4.583** ±0.417	4.125 ±0.528	4.417 ±0.417
Intermediate	**Relevance**	5.000 ±0.000	5.000 ±0.000	5.000 ±0.000	4.875 ±0.311
	Depth	3.667 ±0.807	4.500 ±0.477	3.292 ±0.988	4.708 ±0.396
	Overall	3.958 ±0.450	4.583 ±0.417	3.708 ±0.582	**4.667** ±0.389
Advanced	**Relevance**	5.000 ±0.000	4.833 ±0.389	4.667 ±0.651	4.750 ±0.452
	Depth	3.667 ±0.937	4.583 ±0.469	3.833 ±1.008	4.583 ±0.469
	Overall	4.000 ±0.564	**4.542** ±0.450	3.917 ±0.793	4.500 ±0.369
Combined	**Relevance**	4.944 ±0.199	4.917 ±0.280	4.875 ±0.403	4.833 ±0.359
	Depth	3.708 ±0.740	4.542 ±0.437	3.611 ±0.903	4.500 ±0.521
	Overall	4.021 ±0.457	**4.569** ±0.417	3.917 ±0.649	4.528 ±0.395

Table 1: Evaluation scores for different model configurations (GPT-4o mini and GPT-4o) across introductory, intermediate, and advanced topic complexities. Each metric was scored on a scale of 1 to 5, with 5 representing the best possible score. "Combined" describes the aggregated scores across all complexity levels.

Table 1 displays the human evaluation scores for presentations generated by GPT-4o Mini and GPT-4o models, with and without contextual information, across three levels of topic complexity: introductory, intermediate, and advanced, as well as the combined averaged scores. The evaluation metrics include relevance, depth, and an overall score, providing insights into the quality and comprehensiveness of the information within the generated presentations at each level of complexity.

4.2.1 Presentation Relevance.

The table shows that the relevance scores remain consistently high across all models and topic complexities, with values ranging from 4.667 to 5.000. This consistency indicates that both models are innately proficient at organizing presentations and selecting relevant subtopics for any given topic. The models are particularly effective at ensuring that selected subtopics included in the presentation contribute meaningfully to the overall understanding of the topic. However, the relevance scores after providing context are, on average, slightly lower than their no-context counterparts. This slight decrease could likely be explained due to the models' adherence to the provided context, even when the context focuses on subtopics adjacent to the main presentation topic.

4.2.2 Presentation Depth and Overall Assessment.

The presentation depth, and in turn, the overall presentation scores, were significantly higher when contextual information was provided. For instance, with intermediate-level topics, GPT-4o reached an average depth score of 4.708 with context, a notable increase from 3.708 without context. This trend remained consistent across all factors, with both models exhibiting a combined average depth of around 4.5, compared to around 3.65 in no-context conditions.

Additionally, as the topic complexity increased, the no-context models saw slight performance decreases in overall assessment. The performance of the no-context models slightly declined in both the intermediate and advanced levels, suggesting that these models struggled more with generating more detailed explanations in topics that may not be fully encompassed within the model's parametric knowledge. In contrast, the context models maintained relatively consistent depth and overall scores across all complexity levels, indicating that context plays a crucial role in the robustness of the model. Figure 5 illustrates this contrast, demonstrating two major ways that presentations generated without context can lack depth. The first slide generated by the model without context (top-left) lacks any factual information and instead broadly states "Definition and function of [key term]," while the second (top-right) gives very brief and unspecific details. The bottom slides, generated with context, provide significantly more substantial content, providing clear definitions and explanations of key terms and concepts alongside references to ensure all facts within the presentation are sourceable.

4.2.3 Model Complexity.

We observe a minimal difference in the performances of the GPT-4o Mini and GPT-4o models across all metrics and topic complexities, regardless of whether context was provided. As seen in the "Combined" section of Table 1, GPT-4o Mini with context achieved an overall score of 4.569, only marginally outperforming GPT-4o with context, which scored 4.528. Likewise, without context, both models performed similarly, with the Mini vesion slightly outperforming its larger counterpart.

These results suggest that the GPT-4o Mini model, despite being more compact than GPT-4o, is equally capable of generating high-quality presentations. The presence of contextual information had a far greater influence on the overall performance than the model size, reinforcing the importance of context over computational power when generating detailed presentations.

5 Conclusion

This work has presented a novel pipeline for generating high-quality, customizable presentations that reduce hallucinations by grounding the output with literature, resulting in sourceable information. By integrating RAG, the system ensures that the content is not only relevant, but also verifiable, citing sources from domain-specific corpora to validate generated content. This grounded approach is essential in academic and scientific contexts, where accurate and traceable information is paramount.

This framework is designed to be highly accessible, ensuring that it is not resource-prohibitive for its users. Our results show minimal impact of model size on performance, with the GPT-4o Mini model

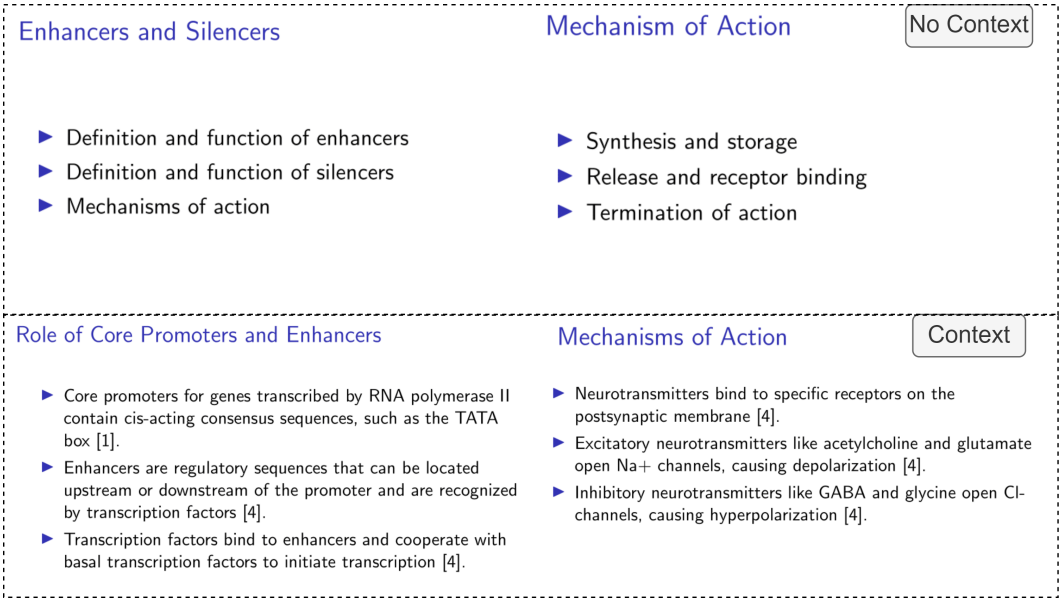

Figure 5: Examples of slides generated by GPT-4o without (top) and with (bottom) context taken from presentations on the intermediate-level topics “Transcription Regulation in Eukaryotes" (left) and “Neurotransmitters" (right). On the bottom-left slide, [1] and [4] refer to *Lippincott Illustrated Reviews: Biochemistry* and *Schwartz's Principles of Surgery*, respectively. On the bottom-right, [4] refers to *Histology: A Text and Atlas : with Correlated Cell and Molecular Biology*. The context-enhanced slides show significant improvements in content depth.

performing comparably to its full-sized counterpart in generating relevant and in-depth presentations, regardless of the topic's complexity. Additionally, the retrieval corpus can easily be swapped with any other collection of texts, regardless of the domain, allowing for flexibility across different fields. Lastly, the system enables collaborative development, allowing instructors to engage with the model to refine and customize the generated slides, adjusting the layout based on the audience's preferences. As AI continues to advance, frameworks like this have the potential to significantly streamline the creation of educational materials, reducing preparation time while ensuring academic rigor and reliability.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Garima Agrawal, Kuntal Pal, Yuli Deng, Huan Liu, and Ying-Chih Chen. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 2024. doi: 10.1609/aaai.v38i21.30362.

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*, 2023.

Michael Alley and Kathryn A Neeley. Rethinking the design of presentation slides: A case for sentence headlines and visual evidence. *Technical communication*, 52(4):417–426, 2005.

B. Alsafari, E. Atwell, A. Walker, and M. Callaghan. Towards effective teaching assistants: From intent-based chatbots to llm-powered teaching assistants. *Natural Language Processing Journal*, 1: 100101, 2024. doi: 10.1016/j.nlp.2024.100101.

Cecilia Arighi, Steven Brenner, and Zhiyong Lu. Large language models (llms) and chatgpt for biomedicine. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*, pages 641–644. World Scientific, 2023.

Salman Bahoo, Marco Cucculelli, Xhoana Goga, and Jasmine Mondolo. Artificial intelligence in finance: a comprehensive review through bibliometric and content analysis. *SN Business & Economics*, 4(2):23, 2024.

Robert A Bartsch and Kristi M Cobern. Effectiveness of powerpoint presentations in lectures. *Computers & education*, 41(1):77–86, 2003.

Patrice Béchard and Orlando Marquez Ayala. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv preprint arXiv:2404.08189*, 2024.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

L. E. Chen, S. Y. Cheng, and J.-S. Heh. Chatbot: A question answering system for students. In *Proceedings of the 2021 International Conference on Advanced Learning Technologies (ICALT)*, pages 345–346. IEEE, 2021.

Francis S Collins, Tara A Schwetz, Lawrence A Tabak, and Eric S Lander. Arpa-h: Accelerating biomedical breakthroughs. *Science*, 373(6551):165–167, 2021.

Conor John Cremin, Sabyasachi Dash, and Xiaofeng Huang. Big data: historic advances and emerging trends in biomedical research. *Current Research in Biotechnology*, 4:138–151, 2022.

Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, Y. Wang, et al. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*, 2023.

Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. How teachers can use large language models and bloom's taxonomy to create educational quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23084–23091, 2024.

O Fagbohun, NP Iduwe, M Abdullahi, A Ifaturoti, and OM Nwanna. Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artifical Intelligence and Machine Learning & Data Science*, 2(1):1–8, 2024.

Tira Nur Fitria. Artificial intelligence (ai) in education: Using ai tools for teaching and learning process. In *Prosiding Seminar Nasional & Call for Paper STIE AAS*, volume 4, pages 134–147, 2021.

Jeffrey S Flier. Publishing biomedical research: a rapidly evolving ecosystem. *Perspectives in Biology and Medicine*, 66(3):358–382, 2023.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. Bioreader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 5770–5793, 2022.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Tanya Gupta. Automatic presentation slide generation using llms. 2023.

Wayne Holmes, Maya Bialik, and Charles Fadel. *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign, 2019.

Yue Hu and Xiaojun Wan. Ppsgen: Learning-based presentation slides generation for academic papers. *IEEE Transactions on Knowledge and Data Engineering*, 27(4):1085–1097, 2015.

S. E. Huber, K. Kiili, S. Nebel, R. M. Ryan, M. Sailer, and M. Ninaus. Leveraging the potential of large language models in education through playful and game-based learning. *Educational Psychology Review*, 36(1):25, 2024. doi: 10.1007/s10648-024-09727-9.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *arXiv preprint arXiv:2401.15269*, 2024.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Robert Leaman, and Zhiyong Lu. Retrieve, summarize, and verify: How will chatgpt impact information seeking from the medical literature? *Journal of the American Society of Nephrology*, pages 10–1681, 2023.

E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, and G. Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023. doi: 10.1016/j.lindif.2023.102274.

Fotis Kitsios, Maria Kamariotou, Aristomenis I Syngelakis, and Michael A Talias. Recent advances of artificial intelligence in healthcare: A systematic literature review. *Applied Sciences*, 13(13): 7479, 2023.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474, 2020.

Richard E Mayer. Multimedia learning. In *Psychology of learning and motivation*, volume 41, pages 85–139. Elsevier, 2002.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023.

Mohammad Muzaffar Mir, Gulzar Muzaffar Mir, Nadeem Tufail Raina, Saba Muzaffar Mir, Sadaf Muzaffar Mir, Elhadi Miskeen, Muffarah Hamid Alharthi, and Mohannad Mohammad S Alamri. Application of artificial intelligence in medical education: current scenario and future perspectives. *Journal of advances in medical education & professionalism*, 11(3):133, 2023.

S. Moore, R. Tong, A. Singh, Z. Liu, X. Hu, Y. Lu, and J. Stamper. Empowering education with llms: The next-gen interface and content generation. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pages 32–37. Springer Nature Switzerland, June 2023.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Wang, and Tom Hope. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, 2022.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Athar Sefid, Jian Wu, Prasenjit Mitra, and C. Lee Giles. Automatic slide generation for scientific papers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Athar Sefid, Jian Wu, Prasenjit Mitra, and Lee Giles. Extractive research slide generation using windowed labeling ranking. In *Proceedings of the Second Workshop on Scholarly Document Processing*. NAACL, 2021. URL `https://doi.org/10.48550/arXiv.2106.03246`.

Kannan Sridharan and Reginald P Sequeira. Artificial intelligence and medical education: application in classroom instruction and student assessment using a pharmacology & therapeutics case study. *BMC Medical Education*, 24(1):431, 2024.

J. Stamper, R. Xiao, and X. Hou. Enhancing llm-based feedback: Insights from intelligent tutoring systems and the learning sciences. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pages 32–43. Springer Nature Switzerland, July 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Mirella Veras, Joseph-Omer Dyer, Morgan Rooney, Paulo Goberlânio Barros Silva, Derek Rutherford, Dahlia Kairy, et al. Usability and efficacy of artificial intelligence chatbots (chatgpt) for health sciences students: protocol for a crossover randomized controlled trial. *JMIR Research Protocols*, 12(1):e51873, 2023.

Sida Wang, Xiaojun Wan, and Shikang Du. Phrase-based presentation slides generation for academic papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Yubo Wang, Xueguang Ma, and Wenhu Chen. Augmenting black-box llms with medical textbooks for clinical question answering. *arXiv preprint arXiv:2309.02233*, 2023.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.findings-acl.372`.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068, 2024.