AIDEN: Automatic Speaker Notes Creation and Navigation for Enhancing Online Learning Experience

Stalin Varanasi[1,2], Umer Butt[1,2], Günter Neumann[1,2], Josef van Genabith[1,2]

[1]Saarland Informatics Campus, D3.2, Saarland University, Germany,
[2]German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany
{stalin.varanasi, mubu01, guenter.neumann, josef.van_genabith}@dfki.de

Abstract

Effective learning in digital environments depends on quick access to educational resources and timely support. We present AIDEN, an advanced, AI-driven virtual teaching assistant integrated into lectures, to provide meaningful support for students. AIDEN's capabilities include reading lecture materials aloud, locating specific slides, automatic speaker notes generation, and search through a video stream. Powered by state-of-the-art information retrieval and text generation, AIDEN can be adapted to new lecture content with minimal manual adjustments, requiring only minor customization of data handling processes and model configurations. Through automated testing, we evaluated AIDEN's performance across key metrics including slide retrieval recall for questions, and alignment of generated speaker notes with ground-truth data. The evaluation underscores AIDEN's potential to significantly enhance learning experiences for offering real-world application and rapid configurability to diverse learning materials.

1 Introduction

1.1 Background

In today's educational landscape, students often face challenges in accessing comprehensive learning materials online and obtaining prompt answers to their questions. Sometimes they do not want to go through a whole lecture but only look for certain subtopics in the lecture to answer a question. Traditional methods of learning sequentially can be time consuming and may not always provide immediate support, which can hinder students' ability to quickly navigate and stay engaged with their coursework. Secondly, speaker notes, which provide relevant information about the lecture slides, are not always provided by the lecturers due to time pressure and/or rapidly changing content. This creates a sparse signal for any on-line chat assistant to navigate through the slides of noncurated courses and provide relevant information to the students.

1.2 Objective

To address these challenges, we developed AIDEN, an interactive chat-based system that serves as a personalized teaching assistant. To enhance this chat system, we designed models that generate speaker notes from presentation slides, implemented information retrieval models to find relevant slides in response to student queries, and created an interface that reads aloud parts of lectures that address those questions while also allowing navigation through online videos. Our system has been tested by students enrolled in two courses in an NLP focused program. Our contributions include:

1. A novel chat interface to support students in any online course featuring presentation slides.
2. An evaluation of Large Language Model based speaker notes generation.
3. Assessment of information retrieval methods for slide retrieval.
4. The development of a speaker notes dataset to train and evaluate chat interfaces for online courses.

1.3 Related Work

Previously, datasets for spoken language in lectures without aligning them with slides have been developed (Cho et al., 2014). With the rise of Large Language Models (LLMs), applications in education have been explored (Geislinger et al., 2022). Additionally, research into AI-generated teaching assistants has emerged (Tack et al., 2023). More recently, a chat interface for courses has been introduced (Sajja et al., 2024). Our work is among

Proceedings of Recent Advances in Natural Language Processing,pages 1299–1303
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-150

the first to tackle various challenges in online learning, such as generating speaker notes and retrieving slides. However, unlike (Sajja et al., 2024), we specifically focus on generating speaker notes from slides and aim to engage students with content approved by lecturers. Recent research has demonstrated the growing potential of educational chatbots in supporting learning processes, with systematic reviews (Guan et al., 2025) and experimental studies (Ait Baha et al., 2024) showing that chatbots can enhance student learning experiences through adaptive interactions, personalized pacing, and improved engagement compared to traditional methods.

2 System Features

AIDEN[1] is a chat interface that helps students navigate course content online. The system provides three core features to support learning:

- **Speaker Notes Generation:** Generates detailed notes from lecture content with additional context and explanations.
- **Content Segmentation:** Segments lecture videos into coherent sections for easier navigation and understanding.
- **Slide Retrieval:** Enables search and retrieval of specific lecture slides during study sessions.

The system supports three main workflows: *Add* reference materials (lectures, textbooks, YouTube videos), *Create* additional materials (slide-to-video conversion, slide extraction), and *Explore* content through the chat interface. Students can interact with all materials through natural language queries, enhancing their learning experience across multiple content formats.

2.1 Add

Add Lectures: Upload and manage lecture slides. Add References: Include textbooks, articles, and research papers. Add YouTube Material: Provide YouTube URLs for indexing and integration.

2.2 Create

Slides to Videos: Convert lecture slides into video presentations with multimedia elements. Videos to Slides: Extract key points from videos and create slides for review.

[1]https://aiden.ngrok.pro/

2.3 Explore

AIden Chat: Interactive interface where students ask questions and receive context-aware answers. Slides: View and navigate uploaded lecture slides. Videos: Watch indexed videos related to lecture topics. YouTube Materials: Access YouTube videos linked to course materials.

3 Question Answering Frame-Work

3.1 Chat Interface

A dialogue system is developed using Google Dialogflow to assist students with their questions. It includes the dialogue acts: *Read Lectures Out*, *Ask Questions from Slides*, and *Ask Questions Online*. The first two intents focus on lecture slides, while the last addresses YouTube videos. These intents facilitate offline lecture listening and question-asking. A text-to-speech audio of the speaker notes is generated using OpenAI models and provided to students on request.

3.2 Speaker Notes Generation

Speaker notes are a critical component of our system, as they provide the additional context and content necessary to support effective question answering for students. To address a lack of speaker notes, we generate synthetic speaker notes by leveraging a large language model (LLM) on slide information. Specifically, we construct prompts using the text from the current slide, the preceding slide, and the subsequent slide to provide contextual continuity[2]. Additionally, we enhance the quality of the generated notes through retrieval-augmented generation (RAG) (Lewis et al., 2020), utilizing a pre-trained retrieval model to incorporate relevant information from external sources. The resulting synthetic speaker notes enrich the lecture material by offering additional context and explanations.

3.2.1 Retrieval Augmented Generation

Knowledge Corpus Construction: We create a domain-specific knowledge corpus using content from an open-source AI resource [3], as our slides are derived from an AI course. The indexing process leverages a pre-trained retrieval model based on FAISS IndexFlatL2 (Johnson et al., 2019). The indexed knowledge base allows efficient similarity searches.

[2]We use gpt-4o-mini-2024-07-18 in our experiments

[3]https://people.engr.tamu.edu/guni/csce421/files/AI_Russell_Norvig.pdf

Retrieval Phase: Given a prompt comprising text from the current, previous, and next slides, we perform a semantic search on the knowledge base using the retrieval model. This step identifies the most relevant passages or documents, ranking them based on similarity scores. We typically retrieve the top-k (e.g., k=5) passages using the Dense Passage Retrieval (DPR) (Kwiatkowski et al., 2019) model to ensure sufficient context. We used the *openai text-embedding-ada-002* model to encode both corpus and query. The query for the retrieval is the text from the slides itself.

Generation Phase: The retrieved passages are then combined with the original input and fed into the generative model, gpt-4o-mini-2024-07-18. Since the input text from slides is often not descriptive enough, the additional retrieved passages provide better context for speaker notes generation. The model conditions its output on both the prompt and the retrieved content, generating comprehensive speaker notes that include insights and explanations beyond the slide content.

3.3 Lecture Segmentation

If a presentation is in video format, we segment the transcriptions to find retrieval candidates for a question. We obtain sentence embeddings using sentence-BERT (Reimers and Gurevych, 2019) and cluster the sentences by identifying relative extremes in the cosine similarities between consecutive sentences. Additionally, we segment by detecting slide changes in the video through frame difference analysis, where we compare consecutive video frames and identify timestamps when the pixel-level differences exceed a predefined threshold, indicating a transition to new slide content.

YouTube Search: We use *whisper*[4] to transcribe and index YouTube videos. We further allow students to search an online YouTube video for the same questions.

3.4 Slide Retrieval

To facilitate effective question answering, we construct a Question-Answer (QA) corpus from the lecture slides. This process leverages Question Generation (QG) techniques applied to speaker notes.

We experimented with two approaches for generating questions:

- **Zero-Shot:** Using GPT-4
- **Supervised Sequence-to-Sequence:** Based on (Varanasi et al., 2020), trained on the Yahoo! Answers Comprehensive Questions and Answers dataset[5].

We generate a training dataset by prompting the large language model on two lectures, as this approach proved more effective at generating contextually relevant and diverse questions for our educational domain. Additionally, we create a Question-Answering dataset containing 31,736 data points from potential reference materials. Our dataset is publicly available[6].

3.5 Infrastructure

For a new online course, the system requires only a set of slides for each lecture. Once the slides are provided, the system automatically generates comprehensive speaker notes and configures the dataset needed to train the slide retrieval model (Karpukhin et al., 2020). This setup process is straightforward and easily adaptable for any new course, ensuring minimal manual effort.

We train our models on a GPU with V100-32GB, leveraging its high computational capabilities for efficient training. During inference, the system runs on an NVIDIA Titan GPU with 12GB of RAM, enabling fast and reliable performance. The dense retrieval model is optimized for speed, indexing 592 text segments from 288 slides in under 2 seconds, even when handling 5 parallel requests. This high retrieval efficiency ensures seamless scalability, allowing the system to handle multiple queries concurrently without compromising performance.

4 Evaluation

To evaluate our system's performance, we create a comprehensive dataset by prompting a Large Language Model (LLM) and partition it into training and testing sets. The dataset is designed for two primary tasks: slide retrieval and speaker notes generation. The evaluation uses a custom dataset with extensively hand-annotated ground-truth speaker notes for four lectures[7]. Our experiments demonstrate the effectiveness of synthetic speaker notes, as shown in Tables 1, 2, and 3.

Dataset Analysis (Table 1): The dataset consists of 1,448 slides in the training set and 288 slides in the test set. On average, the generated speaker

[4]https://openai.com/index/whisper/

[5]https://webscope.sandbox.yahoo.com/catalog.php?datatype=l

[6]https://github.com/StalVars/aiden

[7]*Artificial Intelligence course*, Saarland University

notes in the training set have a length of 161 tokens, while the annotated speaker notes in the test set average 38 tokens. This discrepancy reflects the richness and depth of synthetic notes compared to manually created ones, highlighting the generative model's capability to produce detailed and informative content. The larger training set size ensures robust model training, enabling it to generalize effectively on unseen data.

Dataset	#Questions	#Slides	Avg length
Train	1448	1448	161 (generated)
Test	288	288	38 (annotated)

Table 1: Dataset specifications showing generated vs. annotated speaker notes characteristics

Speaker Notes Generation Performance: To evaluate speaker notes generation, we compare the output from two setups: GPT-4 alone and GPT-4 enhanced with Retrieval-Augmented Generation (RAG) (Table 2). We use BERTScore (Zhang et al.) to measure precision, recall, and F1-score against the ground-truth speaker notes in the test set.

GPT-4 with Retrieval-Augmented Generation (RAG) outperforms the standalone GPT-4 by a slight margin across all metrics, achieving an F1-score of 0.85 compared to 0.84. This small improvement highlights the advantage of retrieval-augmented generation, where incorporating relevant retrieved content helps produce more informative and contextually accurate responses. With high-quality reference material and an enhanced embedding space, performance could potentially improve further.

Both configurations demonstrate strong precision and recall, showcasing the ability of large language models to generate synthetic speaker notes that closely align with human-written ones. The consistent recall across both setups reflects a robust understanding of context, while the increase in precision with RAG indicates enhanced specificity and relevance.

Model	Precision	Recall	F1
GPT-4	0.84	0.85	0.84
GPT-4 + RAG	0.85	0.85	0.85

Table 2: BERTScore evaluation of generated vs. ground-truth speaker notes

Slide Retrieval Performance: For the slide retrieval task, we compare our fine-tuned DPR model against BM25 and a baseline DPR model trained on Natural Questions (NQ) (Kwiatkowski et al., 2019). We evaluate retrieval performance using Recall@k metrics, where k = 5, 10, 20 (Table 3).

Our fine-tuned DPR model outperforms both BM25 and DPR (NQ) across all recall levels, achieving a Recall@20 of 0.90. While BM25 shows strong lexical matching, it falls short on semantic tasks. The lower performance of baseline DPR (NQ) model highlights the value of domain-specific fine-tuning.

These results confirm that the synthetic dataset enhances retrieval performance by enabling specialized model training. The fine-tuned DPR model, leveraging synthetic data, significantly outperforms the baseline and BM25, demonstrating its ability to handle domain-specific content efficiently.

Model	R@5	R@10	R@20
DPR(NQ)[1]	0.43	0.55	0.68
BM25	0.67	0.77	0.85
DPR (ours)	**0.71**	**0.87**	**0.90**

Table 3: Retrieval performance comparison across different models

5 Conclusion

In this work, we introduce a system designed to assist both teachers and students in educational environments by automating the generation of speaker notes and enhancing slide retrieval. By using synthetic speaker notes generated through prompts from a Large Language Model (LLM) and fine-tuning a Dense Passage Retrieval (DPR) model with synthetic data, the system can be seamlessly integrated into online courses, making it a valuable tool for course creation and delivery.

For teachers, AIDEN reduces preparation time by generating speaker notes from slides and aiding slide retrieval. For students, the system provides quick access to specific lecture content through semantic search, with automatically generated notes offering detailed explanations for better comprehension.

6 Acknowledgements

This project was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) as part of the project TRAILS (01IW24005).

[1]Baseline model trained by Kwiatkowski et al. (2019)

References

Tarek Ait Baha, Mohamed El Hajji, Youssef Es-Saady, and Hammou Fadili. 2024. The impact of educational chatbot on student learning experience. *Education and Information Technologies*, 29(8):10153–10176.

Eunah Cho, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2014. A corpus of spontaneous speech in lectures: The kit lecture corpus for spoken language processing and translation. In *LREC*, pages 1554–1559.

Robert Geislinger, Benjamin Milde, and Chris Biemann. 2022. Improved open source automatic subtitling for lecture videos. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 98–103.

Rui Guan, Mladen Raković, Guanliang Chen, and Dragan Gašević. 2025. How educational chatbots support self-regulated learning? a systematic review of the literature. *Education and Information Technologies*, 30:4493–4518.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 3982. Association for Computational Linguistics.

Ramteja Sajja, Yusuf Sermet, Muhammed Cikmaz, David Cwiertny, and Ibrahim Demir. 2024. Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education. *Information*, 15(10):596.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The bea 2023 shared task on generating ai teacher responses in educational dialogues. *arXiv preprint arXiv:2306.06941*.

Stalin Varanasi, Saadullah Amin, and Guenter Neumann. 2020. Copybert: A unified approach to question generation with self-attention. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 25–31.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.