LLMs and NLP for Generalized Learning in AI-Enhanced Educational Videos and Powering Curated Videos with Generative Intelligence

Abstract

The rapid advancement of Large Language Models and Natural Language Processing technologies has opened new frontiers in educational content creation and consumption. This paper explores the intersection of these technologies with instructional videos in computer science education, addressing the crucial aspect of generalization in NLP models within an educational context. With 78% of computer science students utilizing YouTube to supplement traditional learning materials, there's a clear demand for high-quality video content. However, the challenge of finding appropriate resources has led 73% of students to prefer curated video libraries. We propose a novel approach that leverages LLMs and NLP techniques to revolutionize this space, focusing on the ability of these models to generalize across diverse educational content and contexts. Our research utilizes the cubits.ai platform, developed at Princeton University, to demonstrate how generative AI, powered by advanced LLMs, can transform standard video playlists into interactive, AI-enhanced learning experiences. We present a framework for creating AI-generated video summaries, on-demand questions, and in-depth topic explorations. Our approach not only enhances student engagement but also provides a unique opportunity to study how well these models generalize across different educational topics and student needs.

Keywords: *Instructional videos, AI-enhanced learning, Large Language Models (LLMs), Natural Language Processing (NLP), generalization in NLP, computer science education, cubits.ai platform, AI-generated content, interactive video experiences, video summarization, on-demand questions, personalized learning, active learning, data-driven insights, generative AI, educational technology, adaptive learning environments*

1 Introduction

The landscape of computer science education is rapidly evolving, with instructional videos becoming an increasingly integral part of the learning process. Recent surveys indicate that over 78% of computer science students frequently turn to YouTube as a supplementary resource to their textbooks and classroom instruction. This trend underscores a growing demand for high-quality, accessible video content in educational settings.

However, the abundance of online resources presents its own challenges. The task of finding appropriate and reliable content can be overwhelming for students, leading to inefficient learning experiences. This difficulty has resulted in a significant preference shift, with 73% of students expressing a desire for curated video libraries that are tailored to their specific course requirements.

In response to these challenges and preferences, we propose a novel approach that harnesses the power of Large Language Models (LLMs) and Natural Language Processing (NLP) technologies to transform the landscape of educational video content. Our research focuses on the intersection of these advanced AI technologies with instructional videos, particularly addressing the crucial aspect of generalization in NLP models within an educational context.

This paper presents a framework for creating AI-enhanced learning experiences that go beyond traditional video playlists. By leveraging generative AI

powered by advanced LLMs, we demonstrate how standard instructional videos can be transformed into interactive, personalized learning tools. Our approach not only fosters active learning and personalized education but also serves as a testbed for evaluating the adaptability of LLMs across diverse computer science topics and varying student proficiency levels.

2 Background

2.1 The Rise of Video-Based Learning

The proliferation of online video platforms has significantly impacted the educational landscape, particularly in computer science. YouTube, in particular, has become a go-to resource for students seeking to supplement their formal education. The preference for video content stems from its ability to provide visual demonstrations, step-by-step explanations, and the flexibility to pause, rewind, and revisit complex concepts.

2.2 Challenges in Content Curation

Despite the abundance of educational videos, students often struggle to find content that aligns precisely with their course requirements. This challenge has led to a growing demand for curated video libraries, where instructors select and organize relevant content to complement their curriculum.

2.3 The Potential of LLMs and NLP in Education

Large Language Models and Natural Language Processing technologies have shown remarkable capabilities in understanding and generating human-like text. These advancements present an opportunity to enhance the educational video experience by providing personalized summaries, generating relevant questions, and offering in-depth explanations tailored to individual student needs.

2.4 Generalization in NLP for Educational Contexts

A key focus of our research is exploring how well LLMs can generalize across diverse educational content and contexts. This aspect is crucial for developing adaptive, personalized learning environments that can cater to a wide range of topics and student needs within computer science education.

3 Methodology

Our research utilizes the cubits.ai platform, developed at Princeton University, to demonstrate how generative AI can transform standard video playlists into interactive, AI-enhanced learning experiences. The methodology encompasses several key components:

3.1 The cubits.ai Platform

cubits.ai is an innovative platform designed to enhance the impact of computer science courses. It serves as a foundation for our research, providing:

- A comprehensive library of curated video courses tailored to meet academic requirements
- Integration of interactive elements such as embedded quizzes and cuGPT, an intelligent assistant designed to guide students through their learning experience
- Synchronized transcript functionality for efficient content navigation, enabling students to search for specific sections of videos and navigate directly to the relevant parts

The platform's architecture is built around several key components:

1. Content Management System (CMS): A robust system that stores and organizes high-quality videos in a curated content repository. Each video is meticulously tagged with metadata, such as subject, difficulty level, and duration, enabling precise search and filtering. Content curation is managed by subject matter experts, ensuring that only high-value, relevant material is featured.
2. Modular Content Delivery: Videos are divided into segments or chapters, allowing users to access specific parts of a larger video without watching the entire content. This structure

makes it easier for users to focus on particular topics or sections.

3. Personalized User Experience: Users can create individual profiles, receive recommendations based on past views, track their progress, and bookmark specific sections for easy access. The UI and UX are designed to provide a flexible learning environment.

4. Micro-payment Integration: The platform features a flexible micro-payment structure, allowing users to purchase only the content they need, either by paying for individual videos or specific segments. Payments are facilitated through various methods, including digital wallets, credit/debit cards etc, offering users maximum flexibility to engage with the content without the need to buy entire courses.

3.2 Vector Database Implementation

We employ vector databases to structure and query high-dimensional vectors (typically embeddings) that represent data points related to the curated videos. This approach enables more semantically meaningful searches, allowing users to leverage the power of LLMs while remaining focused on their specific learning context.

The process involves:

1. Training a vector database using content from a particular domain, such as computer science. The specific domain content is converted into vector representations using a pre-trained embedding model. This embedding process transforms the content into high-dimensional vectors that encode semantic meaning, where content with similar meaning will have embeddings that are closer to each other in the vector space.

2. Indexing these vectors in a database for fast retrieval. These embeddings represent domain-specific knowledge and are stored in the database for fast retrieval.

3. Ensuring domain relevance by curating the data that goes into the vector database to focus solely on the domain of interest (e.g., computer science, finance, healthcare, etc.). This curation maintains domain relevance and restricts answers to the specific domain.

4. Applying additional filtering using metadata tags that identify sub-domains or contexts (e.g., product category, knowledge type). When querying, we can further restrict the results to specific sub-domains by applying filters based on these metadata tags, ensuring more precise domain-specific outcomes.

When a query comes in, it is converted into an embedding vector using the same model as before. The algorithm then searches the vector database for the closest matches (nearest neighbors) to this query vector. Since the vector database contains only domain-specific vectors, the results will naturally be restricted to that domain.

If needed, the model used to generate embeddings can be fine-tuned on domain-specific data. For example, fine-tuning an embedding model on computer science courses ensures that the embeddings generated from queries are more aligned with the specific language and structure of the computer science domain.

3.3 AI-Generated Content Creation

Our framework facilitates the creation of:

1. Video summaries: Concise overviews of video content, highlighting key points and concepts

2. On-demand questions: Automatically generated questions that test understanding and promote active learning

3. In-depth topic explorations: Detailed explanations and additional resources for students seeking to delve deeper into specific topics

3.4 Integration with Existing Video Platforms

We demonstrate how educators can enhance their existing video playlists by incorporating AI-generated content. This integration focuses on increasing student engagement and establishing safety measures for AI use in education. The ability for any instructor

to curate (for free) a video course makes cubits.ai accessible to all students.

3.5 Generalization Testing

To address the crucial aspect of generalization in NLP models, we implement:

1. Cross-topic evaluation: Testing the model's performance across various computer science subjects

2. Adaptive content generation: Assessing the model's ability to tailor content to different student levels and learning styles

3. Contextual understanding: Evaluating how well the model maintains coherence and relevance across different educational contexts

3.6 Monetization and Access Control

The platform incorporates a monetization layer with a dynamic pricing model, allowing content creators to set prices based on factors such as video length, complexity, or popularity, with the platform taking a commission from each transaction. Prices can be adjusted dynamically based on demand or user engagement.

Users have flexible payment options, including:

- Micro-payments for individual pieces of content

- Subscription plans for bundled access to specific categories of content at discounted rates, offering more cost-effective choices for regular users.

A secure payment gateway supports various payment methods, including credit/debit cards, digital wallets. This secure gateway enables quick and safe payments for users. A token-based access control system grants users access to purchased content, allowing them to revisit and view the content as often as needed without repurchasing, offering a flexible and user-friendly experience.

4 Results and Discussion

4.1 Platform Adoption and User Engagement

cubits.ai has become widely adopted in higher education, with many instructors integrating it into their online learning systems. The platform is already serving thousands of students, providing access to a comprehensive library of curated video courses tailored to meet academic requirements.

Key findings include:

- cubits.ai is consistently ranked as the most useful resource in student surveys conducted since 2020

- The platform's user-friendly interface and well-organized content repository have made it a valuable resource for students

- The integration of interactive elements such as embedded quizzes and cuGPT has enhanced the learning experience

- Students find the platform both engaging and efficient in supporting their learning needs

- The synchronized transcript functionality enables students to search for specific sections of videos and navigate directly to the relevant parts

- By aligning with course materials and offering high-quality video content, cubits.ai has established itself as a vital tool for enhancing educational outcomes

4.2 Impact on Learning Outcomes

While comprehensive studies on long-term learning outcomes are ongoing, preliminary data suggests that AI-enhanced videos are contributing to:

- Increased student engagement with course material

- Improved understanding of complex concepts

- More efficient study practices, with students able to quickly locate and revisit key content

- A more personalized and effective learning experience
- Flexibility in learning, with videos acting as a replacement or engaging supplement to traditional courses

As more students see videos as the primary way to gain knowledge, instead of traditional textbooks, cubits.ai is becoming an important part of the content provided to students.

4.3 Generalization Capabilities of LLMs in Educational Contexts

Our research provides insights into how well LLMs generalize across different educational topics and contexts. We observed that:

- LLMs demonstrate strong capabilities in generating relevant summaries and questions across various computer science topics
- The models' performance can vary depending on the specificity of the subject matter, with more niche topics sometimes requiring additional fine-tuning
- The use of domain-specific vector databases significantly improves the relevance and accuracy of generated content

4.4 Transformative Potential in Large Classes

Drawing insights from computer science courses at Princeton and Rutgers Universities, we highlight the transformative potential of AI-enhanced videos in promoting active learning, particularly in large classes. Key observations include:

- Increased participation and engagement in discussion forums related to video content
- More personalized learning experiences, even in classes with high student-to-instructor ratios
- Improved ability for instructors to identify and address common misconceptions or areas of difficulty

5 Future Work

Future research directions include:

1. Longitudinal studies on the impact of AI-enhanced videos on long-term learning outcomes
2. Exploration of more advanced personalization techniques to tailor content to individual learning styles and preferences
3. Investigation of potential biases in AI-generated content and development of mitigation strategies
4. Expansion of the platform to cover a broader range of academic disciplines beyond computer science
5. Further development of generalization testing methodologies for NLP models in educational contexts
6. Enhancement of the vector database implementation to improve search accuracy and efficiency

6 Conclusion

This research demonstrates the transformative potential of integrating LLMs and NLP technologies into educational video content. By leveraging these advanced AI capabilities, we can create more engaging, interactive, and personalized learning experiences for students. The cubits.ai platform serves as a proof of concept, showcasing how AI-enhanced videos can revolutionize computer science education.

Our findings not only contribute to the ongoing dialogue about generalization in NLP but also provide practical insights into the application of LLMs in educational settings. By bridging these domains, we have established a shared platform for state-of-the-art generalization testing in NLP within an educational framework.

As we continue to refine these technologies and gather more data on their impact, we anticipate that AI-enhanced educational videos will play an increasingly crucial role in the future of higher education. This work serves as a cornerstone for catalyzing research on generalization in the NLP community, particularly focusing on the application and evaluation

of LLMs in adaptive, personalized learning environments.

The scalable architecture of the cubits.ai platform, combined with its micro-payment structure, offers a flexible and accessible approach to high-quality learning. As more students turn to video content for knowledge acquisition, platforms like cubits.ai are poised to become integral components of modern digital education, making high-quality learning both affordable and accessible.

Limitations

While our research demonstrates promising results, it is important to acknowledge several limitations:

1. Domain Specificity: The current implementation focuses primarily on computer science education. Generalization to other academic disciplines may require additional research and model adaptations.

2. Data Privacy Concerns: The use of AI in educational settings raises important questions about data privacy and security, particularly when dealing with student interactions and performance data.

3. Potential for Bias: LLMs trained on large datasets may inadvertently perpetuate biases present in the training data. Ongoing work is needed to identify and mitigate these biases in educational contexts.

4. Scalability Challenges: As the platform grows, there may be technical challenges in scaling the infrastructure to handle increased demand and more diverse content.

5. Limited Long-term Data: While initial results are promising, long-term studies on the impact of AI-enhanced videos on learning outcomes are still in progress.

6. Accessibility Considerations: The current implementation may not fully address the needs of students with disabilities, requiring further work on accessibility features.

These limitations highlight the need for continued research and development in this field, as well as ongoing collaboration with educators and students to refine and improve the technology.

Ethics Statement

This research adheres to the ACL Ethics Policy. We have taken the following ethical considerations into account:

1. Data Privacy: All student data collected through the cubits.ai platform is anonymized and handled in compliance with relevant data protection regulations.

2. Informed Consent: Students are informed about the use of AI in their learning materials and have the option to opt out of data collection for research purposes.

3. Bias Mitigation: We are actively working to identify and mitigate potential biases in the AI-generated content, with a focus on ensuring equitable learning experiences for all students.

4. Transparency: The use of AI-generated content is clearly communicated to students, and we provide explanations of how the technology works to promote understanding and trust.

5. Human Oversight: While leveraging AI technologies, we maintain human oversight in content curation and quality control to ensure the accuracy and appropriateness of educational materials.

6. Accessibility: We are committed to improving the accessibility of our platform to ensure that students with disabilities can benefit from AI-enhanced learning experiences.

7. Environmental Impact: We are mindful of the computational resources required for running LLMs and are exploring ways to optimize our models for energy efficiency.

We are committed to ongoing ethical review and improvement of our research and platform as the field

of AI in education continues to evolve. Regular consultations with ethics boards, educators, and students will be conducted to address emerging ethical concerns and ensure responsible development and deployment of our technology.

Acknowledgements

We would like to express our sincere gratitude to the faculty and students at Princeton University and Rutgers University for their invaluable participation in this research. Their feedback and insights have been crucial in shaping and refining our approach.

Special thanks go to the development team of cubits.ai for their tireless efforts and technical support in bringing this platform to life. Their expertise and dedication have been instrumental in realizing our vision for AI-enhanced educational videos.

We are also grateful to our colleagues in the Computer Science departments at Rutgers University and Princeton University for their constructive feedback and stimulating discussions throughout this research project.

Lastly, we extend our appreciation to the anonymous reviewers whose insightful comments and suggestions have greatly improved the quality of this paper.

7 References

References

[1] *cubits.ai - AI-Powered Learning Platform*. [online] Available at: `https://www.cubits.ai/`.

[2] Ramly, N., Rosli, A.N., Suhaimi, S., Wahab, M.H.A. and Ariffin, A.H., 2020. The Effects of Using Educational Videos in Online Learning: A Case Study for Basic Computer Science Subject. *International Journal of Emerging Technologies in Learning (iJET)*, 15(24), pp.254-266.

[3] Brame, C.J., 2016. Effective educational videos. *Vanderbilt University Center for Teaching*.

[4] Giannakos, M.N., Krogstie, J. and Aalberg, T., 2017. Video-based learning ecosystem to support active learning: application to an introductory computer science course. *Smart Learning Environments*, 4(1), pp.1-13.

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

[6] Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J. and Redmond, P., 2021. A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. *IEEE Access*, 9, pp.102220-102235.

[7] Kgosietsile, T. and Okike, E.U., 2022. An Intelligent Semantic Vector Search Model for Grading and Assessing Students. *International Journal of Advanced Computer Science and Applications*, 13(6), pp.140-151.

[8] Nguyen, A., Ngo, H.N., Hong, Y., Dang, B. and Nguyen, B.P.T., 2022. Ethical principles for artificial intelligence in education. *AI and Ethics*, pp.1-15.

[9] Mohan, G.B., Kumar, R.P., Krishh, P.V., Keerthinathan, A., Lavanya, G., Meghana, M.K.U., Sulthana, S. and Doss, S., 2023. An analysis of large language models: their impact and potential applications. *Journal of Innovation in Computer Science and Engineering*, 12(2), pp.104-115.

[10] Kretzschmar, V., Sailer, A., Wertenauer, M. and Seitz, J., 2024. Enhanced Educational Experiences through Personalized and AI-based Learning. *International Journal on Studies in Education (IJonSE)*, 6(2), pp.191-209.