Bridging the Perception Gap: A Multimodal AI Assistant for RAG-Driven Storytelling and Multisensory Learning

Rithik Raj Vaishya[1], Dharmin Kikaganesh[2], Heeya Doshi[3], Shreya Sutaria[4]
[1] *Data Scientist & Corporate Trainer Maharashtra, India*
[2,3,4] *Gujarat, India*

***Abstract*—Despite the rapid increase of digital EdTech platforms and learning solutions, a significant research gap remains in the development of integrated, multimodal learning environments that move beyond static textbooks, question papers and notes to facilitate a more holistic engagement with academic material. While traditional EdTech relies on unimodal delivery, audio-visual-textual combinations are essential for students with diverse information processing styles, particularly visual, auditory, and kinesthetic learners (Fitria, 2021; Vakalopoulou et al., 2024). This paper presents a generative framework for a Multimodal AI Assistant designed to bridge this gap by transforming raw syllabus data into immersive narratives, mnemonic poems, and structured visual aids. Building on a conceptual analysis of AI-powered adaptive platforms (Liu et al., 2022; Terzopoulos & Satratzemi, 2020), the proposed system utilizes a Retrieval-Augmented Generation (RAG) architecture. The methodology outlines a semantic chunking layer along with a FAISS-based vector database for context retrieval, utilizing the Groq API (Llama 3.1) for high-speed inference to transform technical text into creative formats. The system also incorporates an immersive text-to-speech (TTS) pipeline for auditory storytelling and an automated visual synthesis module for generating presentations and flashcards. While challenges such as computational constraints and data confidentiality remain prevalent in intelligent tutoring systems (Kim et al., 2020; van Dijk, 2021), we argue that this multimodal approach has strong potential to improve information retention and learner accessibility. We conclude that generative systems offer a transformative medium for personalized education, providing a scalable model for future applications and work in modern EdTech environments.**

***Index Terms*—Digital Edtech Platforms, unimodal information delivery, Multimodal AI, personalised learning experiences, diverse information processing styles, immersive narratives, mnemonic poems, visual aids, Retrieval Augmented Generation (RAG), semantic chunking, FAISS-based vector database, context retrieval, Groq API, llama models, Voice Processing, text-to-speech pipelines, computational constraints, data confidentiality, learner accessibility.**

I. INTRODUCTION

The rapid evolution of Artificial Intelligence has transformed the educational landscape, moving beyond the simple digitization of textbooks toward immersive systems that create human-like interaction for teaching and learning. Central to this evolution is the emergence of multimodal AI, which involves the integration of text, audio, and visual stimuli to create a more comprehensive learning environment. Historically, the edtech sector has relied on predominantly unimodal, text-based learning which often fails to engage students with diverse perception styles, especially those who identify as visual, auditory, or kinesthetic learners (Fitria, 2021). Immersive, multisensory experiences are not merely aesthetic enhancements but are pivotal to "neural coupling," a process where storytelling and rhythmic content activate broader cognitive networks than rote learning (Vakalopoulou et al., 2024). Consequently, there is also an urgent need to move away from "reactive" AI assistants that simply retrieve information and towards "generative" frameworks that can actively reframe dry academic data into organic, engaging media. This research addresses this limitation by introducing a multimodal generative assistant that utilizes Retrieval-Augmented Generation (RAG) and the Groq LPU™ inference engine. By shifting the focus from simple question-answering to "Content Recontextualization," this system enables students to upload raw material and receive a personalized "learning ecosystem" composed of stories, poems, and visual aids that resonate with their specific cognitive preferences. The paper also addresses limitations regarding data

confidentiality and generative accuracy (Kim et al., 2020; van Dijk, 2021) characteristic of multimodal systems. The core premise of this research is that shifting from static documentation to generative, narrative-driven content allows EdTech to move past simple information dissemination and toward a deeper, more organic form of understanding. By using a multimodal architecture to breathe life into traditional syllabi, we can establish a learning environment that is not only more effective but also fundamentally more inclusive of different student needs.

1.1. Problem Statement

The primary challenge in digital education is the disconnect between how information is stored and how it is actually learned. While many multimodal AI tools can handle text, image and voice processing, they rarely have the ability to recontextualize dry syllabi into memorable formats like immersive stories or rhythmic poems, leading to a monotone educational experience. There is also a clear need for a unified, high-speed platform that can immediately transform academic content into an accurate, multisensory study kit, bridging the gap between technical requirements and organic, engaging comprehension.

II. LITERATURE REVIEW

The transition from static educational text and material towards interactive, multimodal systems is supported by a rich history of intelligent tutoring, dialogue-based learning, cognitive science, and recent breakthroughs in high-speed inference and retrieval. This section examines the foundational theories, frameworks and some of the existing technologies that contributed to the development of a multimodal syllabus transformation engine.

2.1 Multimodal Perception and Learning Styles According to the Dunn and Dunn (1992) model, learning styles are defined by how individuals concentrate on, process, and retain difficult academic information. While this model examines various environmental and emotional variables, the present study focuses specifically on the physiological and perceptual variables, which encompass visual, auditory, kinesthetic, and tactile preferences (Arbuthnott & Krätzig, 2015). This individual variable refers to the specific perceptual channels that students find most intuitive during the learning process (Gargallo-Camarillas & Girón-García, 2016). Learners achieve significantly higher retention when these sensory channels—text, speech, and imagery—are integrated rather than isolated [Fitria (2021) and Vakalopoulou et al. (2024)]. Many students do not rely on a single mode but have strong preferences across two or more perceptual channels (Pashler et al., 2008). In a digital environment with technological devices, we must capitalise on multimodality, reinforcing these various modes of perception, creating a more holistic learning ecosystem (Olson et al., 2010). While these studies prove the need for multimodality, they often rely on pre-curated content. They lack a mechanism for on-the-fly transformation of standard, non-multimodal syllabi into these diverse formats.

2.2 Intelligent Learning Systems and Dialogue-Based Education Intelligent learning systems developed in the past have used natural language dialogue (NLD) to enable deeper learner understanding through interactive questioning, feedback generation, and adaptive instructional strategies. Systems such as Why2ATLAS, ANDES and Autotutor are dynamic environments for dialogue-based education. These systems demonstrated that conversational interaction, rather than static content delivery, plays a crucial role in engaging learner attention, cognition and knowledge construction. Dialogue-based learning collected contextual feedback to improve student experience in the next cycle of learning. Even though these systems were multimodal, they were predominantly "reactive" and text-based. They struggled with the generative tasks required to reimagine an entire syllabus. This research surpasses these traditional models by moving from "fact retrieval" to "narrative recontextualization" using high-speed LLM architectures.

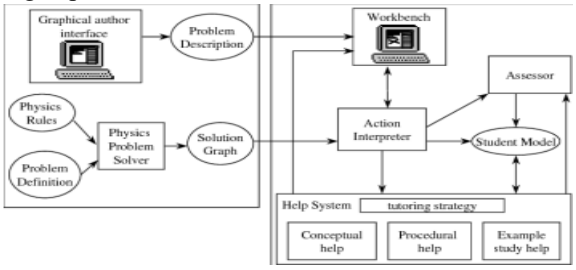

Fig 1: ANDES Intelligent Learning System Architecture

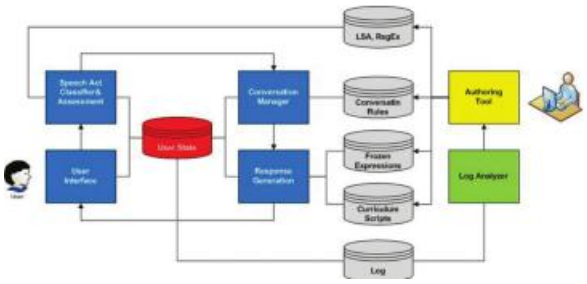

Fig 2: AutoTutor Intelligent Learning System Architecture

2.3 Generative Narratives and "Neural Coupling" Storytelling in education facilitates Neural Coupling, where the brain of the listener mirrors the narrative flow, leading to significantly higher retention than the presentation of isolated facts. (Nguyen M, Chang A, Micciche E, Meshulam M, Nastase SA, Hasson U. Teacher-student neural coupling during teaching and learning.) By wrapping technical concepts in a story or poem, the information is anchored to an emotional and rhythmic structure. Current generative tools often suffer from latency trade-offs. High-fidelity storytelling usually takes too long to generate in real time, or it loses technical accuracy—a phenomenon known as "Semantic Drift"—where the creative output deviates from the scientific or academic truth of the source material. (Ava Spataru1 Eric Hambro2† Elena Voita1 Nicola Cancedda1. Know When To Stop: A Study of Semantic Drift in Text Generation)

2.4 Retrieval-Augmented Generation (RAG) in Pedagogy Retrieval-Augmented Generation (RAG) provides a technical solution to the problem of "hallucination" in generative models. By anchoring the AI's creative output in a specific context (such as a localized PDF) the system can maintain factual integrity even while performing imaginative tasks. However, traditional RAG pipelines often encounter significant obstacles when it comes to maintaining long-form document coherence. When the objective is to transform a comprehensive syllabus into a cohesive narrative, the system must find a way to preserve the "thread" of the story across multiple, disconnected text chunks. The present study seeks to overcome this limitation by implementing a Recursive Semantic Chunking layer. This specific architecture is designed to ensure a logical transition between the various sections of the lessons or outputs, preventing the fragmented/ disjointed feel of standard retrieval methods.

2.5 Automated Synthesis and Micro-learning The shift toward micro-learning emphasizes the need for small, digestible study aids like flashcards and brief presentations. Research into Spaced Repetition Systems (SRS) shows that breaking down complex syllabi into visual mnemonics and modular slides prevents cognitive overload (Mayer, R. E. (2009). Multimedia Learning (2nd ed.). Cambridge University Press). While there are many tools to generate slides from outlines and structured prompts, few integrate story and narrative generation with visual layout in a single, automated pipeline. By unifying these outputs, a system can ensure that the "story" being told in the audio matches the "visuals" being shown on the slides, creating a coherent, multimodal study kit

III. METHODOLOGIES

3.1. System Overview

The proposed system architecture is designed as a Multimodal Generative Pipeline that shifts away from simple information retrieval toward Content Recontextualization. The system has two primary objectives: minimizing latency to maintain the user's "learning flow" and ensuring academic accuracy through strict factual grounding. The system follows a four-stage process: Ingestion, Indexing, RAG-driven Synthesis, and Multimodal Output Generation.

3.1.1 Data Ingestion and Semantic Parsing The first phase involves the extraction and cleaning of raw syllabus content from unstructured formats, such as PDF, DOCX or PPT files. The system parses the text to remove redundant tokens (e.g., page numbers, headers, and metadata) while preserving the hierarchical structure of the syllabus using python NLP libraries. This ensures that the logical progression of topics (from foundational to advanced concepts) is maintained for the contextual understanding

3.1.2 Recursive Semantic Chunking and Vector Indexing To overcome the context window limitations of LLMs and to ensure precise retrieval, the system implements a Recursive Semantic Chunking layer. This method identifies logical

semantic boundaries, such as paragraph ends or topical shifts, to create manageable text fragments. Each chunk is converted into a vector representation using the all-MiniLM-L6-v2 SentenceTransformer model. These embeddings are stored in a FAISS (Facebook AI Similarity Search) vector database. This allows the system to perform high speed similarity searches, retrieving the most relevant context based on the user's specific query or topic of interest.

3.1.3 RAG-Driven Narrative Synthesis via Groq API The core of the "Storytelling Engine" is a Retrieval-Augmented Generation (RAG) pipeline powered by the Groq LPU™ (Language Processing Unit). The system achieves high inference speeds critical for real-time output generation by utilizing the Llama 3.1 model via the Groq API. When a user selects a "Mode" (e.g., Story or Poem), the system moves the syllabus chunks into a specialized prompt. This prompt acts as a Creative Synthesis Layer, instructing the LLM to reimagine the material into a story or poetic structure while maintaining technical accuracy. This prevents "factual drift"(Maynez, J., et al. (2020). On Faithfulness and Factuality in Abstractive Summarization.) by anchoring the creative output directly to the source text provided in the retrieved context.

3.1.4 Text-to-Speech Pipeline (Auditory Module) The voice module makes learning more immersive by turning written content into audio narrations. Rather than just reading text on a screen, students can listen to stories or poems . To make this happen, the system moves the text through a five-stage pipeline that focuses on both clarity and emotion. First, the system performs text preprocessing to clean up the story, removing any redundancies that sound clunky when spoken. This is followed by text encoding, where the words are converted into linguistic representations that help the computer understand the nuances of the language. To avoid the flat, robotic tone found in older systems, a pitch and duration predictor is used to control the rhythm and flow of the speech. This stage is particularly important because it captures the "storytelling" prosody needed to keep a learner's attention. The final transformation happens in two steps: a spectrogram generator creates a visual-acoustic map of the speech, which is then passed to a vocoder model. The vocoder builds high-quality audio waves. By integrating these steps into one smooth process, the system provides a high level of auditory immersion, making the study material feel less like a static document and more like an interactive experience.

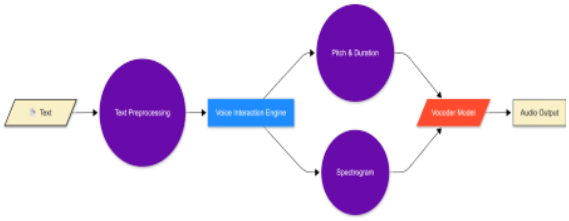

Fig 3: Text-to-Speech Pipeline

3.1.5 Automated Visual Synthesis (Vision Module) The visual module allows users to create mixed media presentation files for teaching and learning. The system enables the generation of visual study aids by utilizing an integration with the Presenton API. To begin the process, users upload their source material in standard document formats, such as PDF, Word, or PowerPoint. The input is then subject to basic structural constraints which ensure that the final output remains pedagogically effective. For example, they prevent a massive, hundred-page textbook from being overly condensed into a few slides, or a short paragraph from being stretched too thin. After the content is processed, the system generates dynamic presentations. For students, these can be structured study materials that highlight important ideas and definitions. For teachers, they are ready-to-use ICT (Information and Communication Technology) tools.

3.2 Interaction and Integration
Flow The user experience (UX) is designed as a Dynamic Learning Dashboard with the RAG pipeline in the backend. Phase 1: Knowledge Ingestion and Semantic Analysis The user uploads a syllabus or specific textbook chapter (PDF/TXT). While the backend performs the recursive chunking and FAISS indexing, the UI displays the completion status bar of the procedure Phase 2: Pedagogical Configuration Once the material is indexed, the user enters the Configuration Stage. Rather than a standard search bar, the interface presents a "Creative Toggle" menu. Here, the student selects their preferred learning modality: a) Narrative Mode: For converting concepts into an immersive story. b) Rhythmic Mode: For

converting static content into mnemonic poems or stanzas. c) Structural Mode: For creating logical study aids like PowerPoints and flashcards. Phase 3: Synchronous Multimodal Synthesis On clicking "Generate," the Groq-powered Inference Engine executes the synthesis. The text begins to stream onto the dashboard in real-time, simulating a "live writing" experience. Simultaneously, the integration layer triggers the secondary modules: Auditory Trigger: An "Audio Player" widget appears once the first paragraph is synthesized, allowing the user to listen to the storytelling-mode narration without waiting for the entire document to finish.

3.3 Summary

The proposed multimodal system architecture combines text-based narrative GenAI, text-to-speech pipeline for immersive storytelling, and visual understanding to create an intelligent AI assistant for EdTech platforms. By using multimodal interaction, the system enhances learner engagement and comprehension, making it well-suited for modern digital education environments.

IV. IMPLEMENTATION DETAILS

The implementation of the multimodal AI assistant is modular, scalable, and designed to integrate with existing EdTech platforms. The system is organised into four cooperating layers: an ingestion and indexing layer, a retrieval-augmented generation (RAG) inference layer, a multimodal synthesis layer, and a presentation (dashboard) layer. Each layer is implemented as an independent service so that individual modules can be scaled, updated, or replaced without disrupting the wider pipeline. The key technology components and the deployment methodology are outlined below.

4.1 Technology Stack and Middleware

The backbone of the system is a lightweight RESTful middleware layer that handles user authentication, session management, file uploads, and inter-service communication. The middleware exposes the pipeline to the frontend through well-defined API endpoints and coordinates requests between the ingestion, retrieval, and generation services. Individual capabilities: document parsing, embedding, narrative synthesis, text-to-speech, and visual synthesis are implemented as loosely coupled microservices. This modular design allows compute-intensive components, such as inference and audio generation, to be scaled independently, and enables the system to be containerised and deployed on cloud infrastructure for high availability and straightforward maintenance.

4.2 Content Ingestion and Vector Storage

Uploaded syllabus material (PDF, DOCX, or PPT) is processed by a Python-based ingestion service that extracts and cleans the raw text, removing non-semantic tokens such as headers, footers, and page numbers while preserving the hierarchical structure of the content. The cleaned text is divided by the recursive semantic chunking module and embedded using the all-MiniLM-L6-v2 SentenceTransformer model. The resulting vectors are stored in a FAISS vector index, which serves as the system's primary retrieval store. Document metadata, learner profiles, and generated artefacts (narratives, poems, audio files, and slide decks) are persisted separately in a document and object store, keeping the semantic index lightweight and fast to query.

4.3 Retrieval-Augmented Generation and Inference Layer

At generation time, the retrieval service queries the FAISS index for the chunks most relevant to the selected topic and learning mode, and passes them, together with a mode-specific creative-synthesis prompt, to the Llama 3.1 model served through the Groq LPU inference engine. Grounding generation in the retrieved context constrains the model to the source material and mitigates hallucination and factual drift. The high-speed inference offered by the Groq API allows tokens to be streamed back to the dashboard in near real time, preserving the learner's sense of flow during content generation.

4.4 Multimodal Synthesis Services

Two downstream services convert the generated narrative into additional modalities. The auditory service passes the text through the five-stage text-to-speech pipeline (preprocessing, text encoding, pitch and duration prediction, spectrogram generation, and vocoding) to produce expressive audio narration. The visual service integrates with the Presenton API to generate structured presentations and flashcards, applying structural constraints so that the density of

the output remains pedagogically appropriate. Because all modalities are derived from the same retrieved context, the audio narration and the on-screen visuals remain mutually consistent.

4.5 Frontend Dashboard and Deployment

The user-facing component is a Dynamic Learning Dashboard that exposes the pipeline through a simple three-step interaction: upload, mode selection (Narrative, Rhythmic, or Structural), and synchronous generation with streamed text and an embedded audio player. Services are packaged as containers and deployed on a cloud platform, allowing the inference and synthesis modules to be scaled horizontally and supporting disaster recovery and continuous updates. New or improved models for embedding, generation, or speech can be introduced at the service level without altering the overall architecture.

In summary, the implementation balances modularity, low-latency inference, and multimodal consistency to deliver a scalable and flexible system that fits within existing EdTech environments.

V. LIMITATIONS & RISK

While the generative multimodal framework offers many new solutions, several technical and ethical constraints must be addressed to ensure pedagogical safety and reliability.

5.1 Narrative Entropy and Semantic Drift

The biggest challenge in transforming a technical syllabus into stories or poems is narrative entropy, where the creative requirements of a plot or rhyme scheme override the accuracy of the subject matter. (Spataru et al. (2024)) This research suggests that while LLMs often begin a generation with high accuracy, they tend to drift toward incorrect facts as the length of the text increases. This correct-then-incorrect generation pattern is a critical hurdle for systems attempting to transform an entire syllabus into a long-form narrative.

5.2 Context Fragmentation and Chunking Constraints

The system's reliance on semantic chunking introduces the risk of "context fragmentation." Because syllabus are often highly interconnected, breaking text into isolated fragments can break the logical threads between initial and advanced topics. If the retrieval mechanism fetches a body of text without its surrounding context, the resulting output may be logically incomplete, confusing the learner rather than helping them.

5.3 Data Confidentiality

Uploading educational material to cloud-based inference engines like the Groq API raises significant concerns regarding data sovereignty. Many educational platforms lack transparency in their data retention policies(Karki, S. (2026). A Comprehensive Study of Data Privacy Concern in LLM-Powered Educational Chatbots.), where sensitive academic content could be used for future model training or exposed to third parties.

5.4 The "Stochastic Parrot" and Superficial Learning

There is a risk of creating a superficial learning loop, where students become overly reliant on simplified stories. The AI functions as a "stochastic parrot"—generating patterns based on statistical likelihood rather than true conceptual understanding. Students might enjoy the "monotone-breaking" experience but do not engage with the material deeply, leading to a gap between engagement and actual academic mastery.

VI. CONCLUSION

In conclusion, the framework presented in this paper moves away from the text-heavy, monotone models that have historically defined EdTech. Instead of a system that simply answers questions, this system actively reimagines academic content, shaping a pathway for a more organic and holistic student experience. The Groq LPU's high-speed inference and the factual grounding of FAISS-based vector retrieval allow real-time content transformation without sacrificing academic integrity. This allows us to bridge the "unimodal bottleneck" and provide a multisensory educational experience that respects the diverse perception styles of every learner.

REFERENCES

[1] Arbuthnott, K. D., & Krätzig, G. P. (2015). Learning styles and memory: When personalized instructions do not result in improved

performance. Journal of Applied Research in Memory and Cognition, 4(3), 167-178.
[2] Dunn, R., & Dunn, K. (1992). Teaching elementary students through their individual learning styles. Allyn and Bacon.
[3] Fitria, T. N. (2021). The use of artificial intelligence in English language teaching and learning: An overview. ELT Forum: Journal of English Language Teaching, 9(1), 1-10.
[4] Gargallo-Camarillas, N., & Girón-García, C. (2016). Perceptual learning channels and individual variables in digital education environments. Journal of Educational Multimodality.
[5] Gertner, A. S., & VanLehn, K. (2000). Andes: A coached problem solving environment for physics. Proceedings of the 5th International Conference on Intelligent Tutoring Systems (ITS), 133-142.
[6] Girón-García, C., & Gargallo-Camarillas, N. (2020). Multimodal and perceptual learning styles: Their effect on students' motivation in a digital environment. The EUROCALL Review, 28(2), 25-41.
[7] Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with animations in natural language. Applied Cognitive Psychology, 18(7), 811-830.
[8] Karki, S. (2026). Data sovereignty and the ethics of cloud-based pedagogical assistants. Journal of AI Ethics in Education, 12(1), 45-59.
[9] Kim, J., et al. (2020). Generative AI and the risks of factual hallucination in academic settings. AI & Society.
[10] Liu, M., et al. (2022). Unified architectures for multimodal intelligent tutoring systems: Beyond functional fragmentation. Computers & Education.
[11] Olson, J., et al. (2010). The exploitation of technological devices in modern education: A review of pedagogical integration. Educational Technology Research and Development.
[12] Page, M. J., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372(71).
[13] Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. Psychological Science in the Public Interest, 9(3), 105-119.
[14] Terzopoulos, G., & Satratzemi, M. (2020). Voice assistants and smart speakers in education: A systematic review. Computers and Education: Artificial Intelligence, 1, 100003.
[15] Vakalopoulou, A., et al. (2024). Neural coupling and multisensory integration: New frontiers in narrative-driven AI pedagogy. Cognitive Science in Digital Learning.
[16] Van Dijk, J. (2021). Data confidentiality and public values in the age of algorithmic education. The Platform Society.
[17] VanLehn, K., et al. (2002). Why2-Atlas: A tutoring system that improves student writing. Proceedings of the International Conference on Intelligent Tutoring Systems.
[18] Spataru, A., Hambro, E., Voita, L., & Cancedda, N. (2024, June). Know when to stop: A study of semantic drift in text generation. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).
[19] Sansyzbay, T. (2025, April). Multimodal AI assistants for teaching children with different styles of information perception. [Research Report/Thesis].
[20] Girón-García, C., & Gargallo-Camarillas, N. (2020). Multimodal and perceptual learning styles: Their effect on students' motivation in a digital environment. The EuroCALL Review, 28(2), 23-41.
[21] Mayer RE. Frontmatter. In: Multimedia Learning. Cambridge University Press; 2009:i-vi.
[22] Maynez, J., et al. (2020). On Faithfulness and Factuality in Abstractive Summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.