AI-Driven Video Learning Framework Using OCR and Generative Language Models

Vaishavi R[1], Ramkishore E[2],
Dr. N.V.S. Sree Rathna Lakshmi[3]
Department of Electronics and Communication Engineering,
SRM Institute of Science and Technology, Ramapuram Campus, Chennai.

Abstract*— Traditional textbook learning is becoming increasingly incomprehensible for students, and the need for intuitive, more engaging learning tools has never been felt more. This study describes the CLARITY platform, which is powered by AI that converts academic content in images, PDFs, or any form of typed content into interactive animated video explanations. The system combines Optical Character Recognition (OCR), Natural Language Understanding (NLU), and AI video synthesization to reduce complexity to digestible, visually rich storytelling. Utilizing Tesseract OCR for text extraction and Google Gemini for semantic explanation generation, narrative scripts are created, structured and automatically transformed into animated video lessons with synchronized audio. It makes concepts easy and quick to grasp and retain for students at all levels-from early school learners to competitive exam aspirants. This bridges the gap between traditional learning and modern digital pedagogy in education, making it more accessible, interactive, and engaging.** ***Index Terms- **AI in Education, Optical Character Recognition (OCR), Gemini API, FastAPI, Streamlit, Android Studio, Automated Video Generation, Natural Language Processing (NLP), Machine Learning, Text-to-Video Conversion, E-Learning, Interactive Learning, Deep Learning, Image Processing.**

INTRODUCTION

The modern education system requires students to understand and remember every minute detail of the concepts and ideas presented in conventional textbooks, which is often a cumbersome and monotonous process. Since text-based educational content is static, it restricts visualization, and therefore, learners frequently fail to conceptualize abstract, historical, or scientific information. This issue is particularly evident among school students and competitive examination candidates, such as TNPSC, NEET, and UPSC aspirants, who are required to study vast volumes of theoretical material with limited opportunities for interactive engagement. Consequently, poor conceptual understanding, reduced attention spans, and increased learning fatigue have emerged as major challenges, primarily due to the unavailability of adaptive, interactive, and engaging study resources. To address these issues, the current research proposes an AI-driven educational framework named CLARITY, designed to automatically convert textbook content into visually engaging animated video lessons. The proposed system integrates advanced technologies such as Optical Character Recognition (OCR), Natural Language Processing (NLP), and Generative Artificial Intelligence to transform static educational content into interactive multimedia explanations. The architecture of CLARITY unifies several components into a seamless pipeline, allowing users to upload text, images, or PDF documents via a mobile or web interface. The system utilizes the Tesseract OCR engine to accurately extract text from scanned materials, which is subsequently processed using Google's Gemini API — a large multimodal generative model that contextualizes the extracted data into rich, explanatory narration optimized for student comprehension. Further processing for synchronized audio is achieved through gTTS, while animated visualizations aligned with the narration are created using MoviePy and AI-based video generation platforms such as RunwayML or Pika Labs. The application backend is developed using FastAPI and tested on Uvicorn to enable real-time asynchronous processing, whereas the frontend validation and prototyping are implemented using Streamlit for efficient visualization. Finally, the mobile version is developed using Flutter in Android Studio, providing an intuitive and user-friendly interface through which students can seamlessly learn via animated videos derived directly from their own study materials. By automating the conversion of static textual content into dynamic multimedia-based lessons, CLARITY effectively bridges the gap between conventional reading and experiential learning, thereby enhancing engagement, comprehension, and long-term retention across all levels of education.

I. Related works

Alahmadi and Alshangiti [11] proved that the optimization of OCR with the help of image super-resolution combined with large language models significantly improves the quality of recognition in video-based learning. A study they conducted in Mathematics shows how the enhanced OCR pipelines extract accurate textual information from low-quality educational videos, becoming an important step in turning textbooks into digital learning content. Similarly, Hukkeri et al. [12], while assessing machine learning-based OCR systems for slide-to-text conversion in lecture videos, concluded that hybrid methods involving CNN-based detection together with transformer-based text correction offer better accuracy in educational applications.

Fei et al. [13] investigated whether contemporary video language models inherently possess OCR capabilities and noted that, while

the LLMs will be able to infer the visual context from the frames, dedicated OCR preprocessing is still very important to maintain text fidelity within video learning environments. Further developing this topic, Mishra et al. [14] proposed a technique for AI-based automation of online course trailer generation. In this approach, natural language understanding and content summarization are in use in order to create small teases about courses in an interesting manner, which is one of the basic concepts of dynamic visualization in AI-powered study aids.

Kallamadugu et al. [15] emphasized the AI-assisted narration workflow for adopting existing voice synthesis systems to actively involve students in learning through audio. The results demonstrated that natural-sounding AI narration can effectively replace human voiceovers, thus improving access to courses for students of varied learning preferences. In this context, Cukurova [16] proposed a synthetic virtual instructor with an ability for adaptive response to learner input-one of the major features of interactive AI tutoring systems devised for personalization and motivation in digital classrooms.

In parallel to these developments on narration and interactivity, Singh presents an extensive survey on AI text-to-image and text-to-video generators, focusing on how textual explanations can be transformed into illustrative visual sequences via diffusion-based architectures. Further to that, Singer et al. proposed the Make-A-Video framework: a state-of-the-art model to synthesize realistic motion videos from textual descriptions without the use of text-video pair data. This paradigm reveals exactly how the transformation of learning subjects-from historical narratives to scientific processes-into vivid animation provides a better understanding.

Further advances in video synthesis came with Saharia et al., where they introduce Imagen Video, a diffusion-based model which enables high-fidelity semantically accurate animation from mere textual prompts. On the other hand, Wu et al. have proposed CogVideo: a large-scale pre-training framework for text-conditioned video generation that proves multimodal transformers can effectively translate linguistic content into coherent animated visual sequences. Both are bound to transform education in which concepts from textbooks or handwritten notes will be dynamically visualized for better learning outcomes. Lewis et al. further developed the retrieval-augmented generation model, RAG, which grounds large language model outputs in relevant contextual data at inference time to reduce the number of hallucinations and improve factual accuracy. Reimers and Gurevych, on the other hand, presented Sentence-BERT, a model deliberately designed to support meaningful embeddings for semantic search and knowledge retrieval fundamental techniques in mapping textual concepts to multimedia generation. Again, Johnson et al. explained large-scale vector similarity methods via FAISS, enabling efficient retrieval from educational databases, while Devlin et al. set the stage with BERT, the basis of most modern transformers providing intelligent reasoning and summarization in intelligent tutoring systems. These works cumulatively present a solid backbone for the proposed AI Study Aid App, as it integrates OCR-based text recognition, AI-driven narration and summarization, and text-to-video generation all unified through retrieval-augmented and embedding-based reasoning. This merges computer vision, natural language understanding, and generative AI into a new paradigm for automated education technology that will allow learners of all ages to visualize and understand complex subjects through intuitive, immersive digital storytelling.

II. METHODOLOGY

CLARITY is a proposed system designed as a modular, asynchronous pipeline that automatically converts image-, PDF-, or text-based educational materials into animated instructional videos. The architecture integrates the following five major components: an input acquisition interface, an OCR preprocessing and text extraction engine, an LLM-based semantic processing module, an AI-driven video synthesis framework, and a backend–frontend orchestration layer. It will enable scalable, low-latency, and secure multi-stage processing to support a variety of learners using diverse formats of content. Rapid prototyping is enabled by Streamlit, while the mobile application user interface is production-ready thanks to Flutter. Asynchronous backend execution is provided by FastAPI with Uvicorn, while dynamic multimedia generation is enabled by Google Gemini, gTTS, MoviePy, and other cloud-based APIs for video creation.

A . Development Environment and Tools

The code is written in Python 3.11, and its suite of libraries includes: FastAPI as the backend framework, Uvicorn for ASGI-based deployment, pytesseract, Pillow, and OpenCV for OCR preprocessing and image enhancement. It leverages gTTS for text-to-speech synthesis and MoviePy to assemble the videos locally. Advanced animated visual generation might use other cloud-based generative video tools like RunwayML or Pika Labs. Streamlit is used during the course of development to quickly validate the results of OCR, Gemini, and video generation. The final mobile app will be developed on Flutter in Android Studio. The back-end will be deployed on Render; secret keys will be kept secure via environment variables. GitHub is used for version control with automated CI/CD.

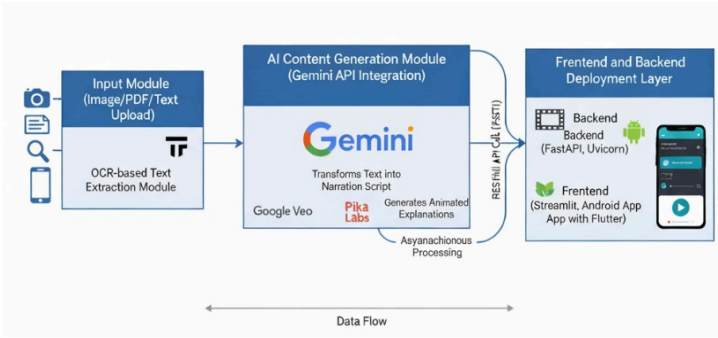

Figure 1 : Block Diagram of Data flow

B . Input Acquisition & Back-end Reception

The first step involves uploading images, PDFs, or simply typed text through the Streamlit prototype or a Flutter mobile application. These are sent to the FastAPI backend through a secure REST endpoint. Upon receiving a request, the backend assigns a unique identifier to the job and temporarily saves the uploaded file for further processing. The asynchronous system will enable the backend to respond to the client while continuing to undertake the other processing tasks asynchronously in the

background and continuously update the user of any progress.

C .OCR Preprocessing and Text Extraction

The pipeline initiates with the conversion of uploaded documents to processable forms, such as images. It includes document rasterization using pdf2image and image preprocessing by OpenCV. Preprocessing involves reduction of noise in the images, gray scaling of the images, use of adaptive thresholding, morphological operations, and super-resolution to enhance the recognition quality. Using Tesseract OCR, text is extracted from the preprocessed images using an appropriate mode of page segmentation. Both the raw text and structural metadata are captured for further refinement. A later stage post-processes this via a rule-based cleaning and LLM-assisted correction into clean, semantically correct text suitable for downstream modeling. E. Semantic Processing Using Google Gemini Given the fine-tuned text input, Google's Gemini large multimodal language model will interpret the extracted academic content and create an educational script of scene descriptions, conceptual explanations, and suggestions of visual elements. In this phase, static textual information is transformed into a pedagogically optimized narrative, fit for animated video lessons. Subtitles, learning objectives, and optional quiz questions also get generated for added instructional value. Figure 2 shows how different preprocessing methods affect the accuracy of OCR, highlighting that enhanced image preparation leads to more reliable text extraction.

Method	Preprocessing Technique	Accuracy (%)	Remarks
Raw Tesseract	No	82.7	Sensitive to noise
Thresholding	Otsu Binarization	88.9	Better for clear prints
Morphological	erosion-dilation	91.5	enhances the feature boundary.
Combined Approach	Denoise + Morph+ Threshold	96.3	Optimal for study material

Figure 2 : OCR Accuracy under Different Preprocessing Methods

D .Semantic Processing Using Google Gemini

With the given fine-tuned text input, Google's Gemini large multimodal language model will interpret the extracted academic content and create an educational script of scene descriptions, conceptual explanations, and suggestions of visual elements. In this phase, static textual information is transformed into a pedagogically optimized narrative, fit for animated video lessons. Subtitles, learning objectives, and optional quiz questions also get generated for added instructional value. Figure 3 demonstrates an intermediate text-processing stage of the CLARITY system, where user-input content is organized into scene-based educational narration through a Gemini-powered semantic model.

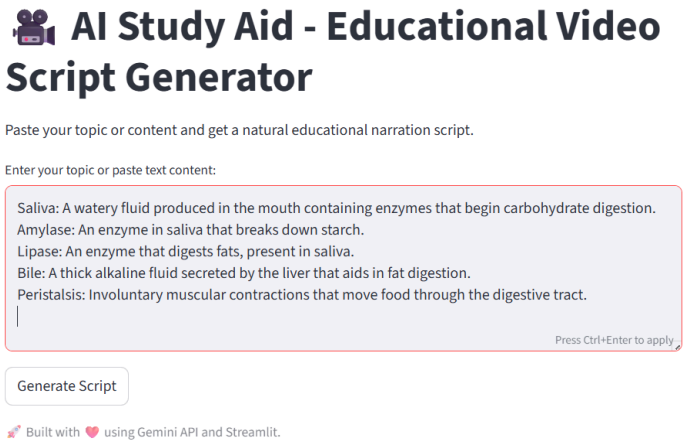

Figure 3 :Streamlit-based prototype of the AI Study Aid script

E . Audio Generation and Video Synthesis

Audio is synthesized, either with gTTS or cloud-based TTS models, to maintain a natural pace and clarity. There are two ways in which video generation proceeds: by local assembly, MoviePy constructs animated slides that include background visuals, text overlays, transitions, and synchronized audio; or through cloud-based video generation via APIs from RunwayML or Pika Labs, dynamic animated scenes are built based on the visual descriptions of Gemini. Segments of video, generated by either means, are stitched together, subtitled, and encoded into MP4 streams or downloadable files.

F . Storage, Delivery, and Frontend Integration

The final video is stored, along with the extracted text and the script metadata, in cloud storage at a location associated with the Render backend. Results are pulled by the client application via REST endpoints where the status of the job keeps updating. During development, the Streamlit interface displays intermediate outputs like extracted text, Gemini narration, and generated videos. The Flutter application provides production-grade controls for playback, bookmarking, subtitle toggling, and options to download files. Figure 4 depicts the transformation of raw text into an AI-generated narrative prepared for animated educational video synthesis.

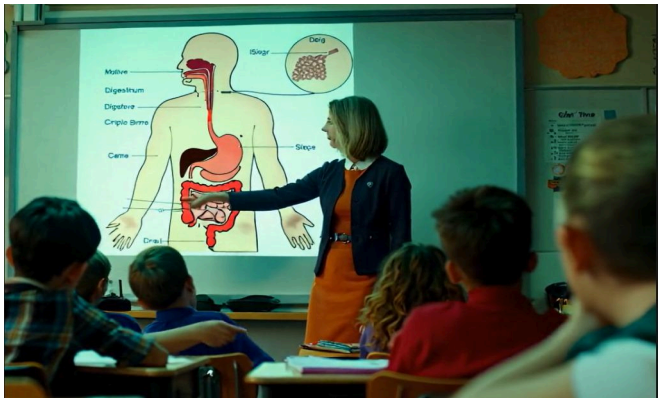

Figure 4 : Traditional Classroom using Static diagrams

G. Security, Privacy, and Compliance

The data in transit is secured by HTTPS, and all sensitive credentials like API keys are stored only in environment variables. Optional AES-256 encryption for user content at rest, along with auto-file-deletion policies, helps comply with GDPR and Indian data privacy requirements. The system processes only user-uploaded or licensed educational content, and it does so in such a way as to guarantee ethical consideration and respect of copyright.

H. Testing, Evaluation, and Performance Metrics

Several performance metrics are considered for this system. CER and WER are the metrics to evaluate the accuracy of OCR. Narration quality is evaluated by human evaluators based on clarity, coherence, and correctness. In the case of videos, the engagement comes through playback analytics, while learning effectiveness is validated through controlled A/B testing. The latency, scalability, and throughput under varying loads should be tested to assure real-time operability.

Figure 5 : CLARITY mobile application welcome screen

Figure 5 is the CLARITY application home screen, where the user can either sign up or log in to access AI-powered educational tools developed on this system. The interface is made simple and easy to use to make it accessible for people from all walks of life.

IV. Future Works

Although the proposed system of CLARITY goes a long way toward automating this educational transformation of text to video, scalability, inclusiveness, adaptability, and pedagogical enhancement are all dimensions for which much potential remains. The system will be further enhanced as an end-to-end intelligent, multilingual, personalized learning environment serving diverse academic needs from rural school learners to higher-education candidates. A few major enhancements envisioned in upcoming releases are discussed below..

A. Multilingual OCR and Generative Narration

Presently, most features support only English-based OCR and narration. In further enhancements, a powerful multilingual pipeline will be provided for major Indian languages, including Tamil, Hindi, Telugu, Malayalam, Kannada, Bengali, and Marathi, including other global languages. This would include training language-specific OCR models, integrating multilingual LLMs, and generating culturally appropriate video narrations. A feature like this will go a long way in making the system accessible to rural students and first-generation learners who use only the vernacular medium.

B. Rural-friendly User Interface and Ultra-Simplified App Flow

Most of the learners from rural and low-digital-literacy regions face problems because of highly complicated interfaces of the apps. Minimalistic and highly intuitive UX design, targeted toward users uninitiated with smartphones, will be incorporated into future versions. Scanning with one tap, voice commands, guided onboarding, and icon-based navigation will ensure usability by even children or older parents. Text-to-voice accessibility, large icons, and offline storage will further enhance usability.

C. Advanced Animation, Interactive Learning, and Storytelling

While animated content is already produced within the present system, sophisticated motion graphics, character-based storytelling, and interactive visual capabilities to be developed will increase learner engagement significantly. Realistic educational animations will reenact historical events, demonstrate scientific processes, and conceptual models in 3D. Scene-based user interactions, adaptive pausing, and micro-visualizations of reinforcing concepts will provide an immersive learning experience for users.

D. Adaptive AI Models for Personalized Learning

Future enhancements will also use the adaptive learning systems that analyze user performance, reading speed, past errors, and knowledge gaps. Based on these patterns, the system intelligently adjusts the length, detail, and complexity of the generated video explanations. Automatic personalized revision sheets, concept maps, and reinforcement questions will also be generated. This adaptive modeling will help students with poor conceptual clarity gain confidence through customized tutoring.

E. AI-powered educational avatars with human-like instruction

CLARITY will bring in AI-driven education avatars able to convey information with expression, gesture, and teacher-like mannerisms for enhanced immersion. These will be set to speak

in regional languages, mimic teaching styles, or adjust emotive tone for better student motivation. The idea behind such multimodal virtual instructors is to help bridge the gap between traditional tutoring and autonomous, AI-powered instruction..

F. Collaborative Learning, Teacher Tools, and Analytics

Future versions will include collaborative classrooms, teacher dashboards, and parent monitoring in order to extend the system beyond individual learning. It also allows educators to upload partial curriculum parts, creates lesson videos, and assigns AI-generated quizzes. Progress can be tracked by both parents and teachers; interventions can be tailored by real-time learning analytics on concept mastery scores, attention metrics, and predictions over learning curves.

G. Integration with National Educational Platforms

The system shall be scaled up long-term by establishing interoperability with the government's digital learning frameworks like DIKSHA, NPTEL, SWAYAM, and NCERT digital resources. This would allow CLARITY to automatically map the generated videos against curriculum standards that can then be deployed at scale across schools and skill-development centers.

V. Conclusion

The CLARITY system provides an end-to-end, intelligent pipeline that transforms traditional textbook content into engaging animated video lessons through integrated OCR, generative AI, and multimedia synthesis. Right from the time a learner opens the CLARITY mobile application, the system allows the user to upload images, PDFs, or text. Further, the Tesseract-based OCR engine accurately extracts information from low-quality or handwritten sources. This content is then understood and semantically enhanced by Google's Gemini multimodal model, which creates a well-structured educational script targeted at instructional clarity, including scene narratives and concept-driven explanations. The synthesized narration is passed to the video generation module, which automatically composes synchronized audio, animated visuals, and explanatory elements into a coherent learning video. The user can instantly preview, download, and study the generated content, reducing the time and expertise required to develop educational material. All in all, this system holds immense promise for improving comprehension, facilitating accessible learning, and catering to a wide range of students, including early learners to competitive exam candidates. With its fast processing, user-friendly interface, and AI-driven pedagogical adaptability, CLARITY represents a promising step toward personalized, multilingual, and immersive digital learning experiences. Figure 6 presents a relative assessment of conventional manual teaching videos and recorded LMS platforms compared to the proposed system, StudyAid. In comparison, whereas the former methods would require several hours of manual effort with very limited adaptiveness, the proposed StudyAid would have a highly automated AI-driven workflow for generating personalized education content within less than five minutes. This illustrates that there is great efficiency improvement in scaling and engagement through the proposed framework.

System	Automation Level	Adaptivity	Content Generation Time	Engennengating
Manual Teaching Videos	Low	Fixed	3–6 hours	Moderate
Recorded LMS (BYJU's, Khan Academy)	Medium	Partial	30–60 minutes	High
StudyAid (Proposed System)	High	Dynamic (AI-based)	<5 minutes	Very High

Figure 6 :Comparative Study with Existing Educational Systems

III. References

[1] 1. R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. Ninth Int. Conf. on Document Analysis and Recognition (ICDAR), 2007. [Online]. Available: https://ai.stanford.edu/~btaskar/ICDAR2007/papers/Smith_Tesseract.pdf

[2] 2. M. D. Alahmadi and M. Alshangiti, "Optimizing OCR Performance for Programming Videos: The Role of Image Super-Resolution and Large Language Models," Mathematics, vol. 12, no. 7, p. 1036, 2024. [Online]. Available: https://doi.org/10.3390/math12071036

[3] 3. G. S. Hukkeri, R. H. Goudar, P. Janagond, and P. S. Patil, "Machine Learning in OCR Technology: Performance Analysis of Different OCR Methods for Slide-to-Text Conversion in Lecture Videos," International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, pp. 325–332, 2022. [Online]. Available: https://doi.org/10.14569/IJACSA.2022.0130839

[4] 4. Y. Fei, Y. Gao, X. Xian, X. Zhang, T. Wu, and W. Chen, "Do Current Video LLMs Have Strong OCR Abilities? A Preliminary Study," in Proc. 31st Int. Conf. on Computational Linguistics (COLING), 2025. [Online]. Available: https://aclanthology.org/2025.coling-main.659.pdf

[5] 5. P. Mishra, C. Diwan, S. Srinivasa, and G. Srinivasaraghavan, "AI-Based Approach to Trailer Generation for Online Educational Courses," arXiv preprint arXiv:2301.03957, 2023. [Online]. Available: https://arxiv.org/abs/2301.03957

[6] 6. A. H. Kallamadugu, N. S. Lawal, and J. M. Burgett, "A Workflow for Creating Narration for Voice-Over Presentation Using Commercially Available Artificial Intelligence," J. ATE Open Access Publishing, 2024. [Online]. Available: https://micronanoeducation.org/wp-content/uploads/2024/08/J-ATE-3-2-A-Workflow-for-Creating-Narration.pdf

[7] 1. M. Cukurova, "Investigating the Potential of Learning Videos with Synthetic Virtual Instructors," British Journal of Educational Technology, forthcoming/2024. [Online]. Available: https://doi.org/10.1111/bjet.13530

[8] 2. A. Singh, "A Survey of AI Text-to-Image and AI Text-to-Video Generators," arXiv preprint arXiv:2311.06329, 2023. [Online]. Available: https://arxiv.org/abs/2311.06329

[9] 3. U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-A-Video: Text-to-Video Generation without Text-Video Data," arXiv preprint arXiv:2209.14792, 2022. [Online]. Available: https://arxiv.org/abs/2209.14792

[10] 4. S. Saharia, J. Ho, W. Chan, D. Fleet, and M. Norouzi, "Imagen Video: High Fidelity Video Generation with Diffusion Models," arXiv preprint arXiv:2305.16709, 2023. [Online]. Available: https://arxiv.org/abs/2305.16709

[11] 5. Z. Z. Wu, M. Yang, H. Guo, and Y. Gao, "CogVideo: Large-Scale Pretraining for Text-to-Video Generation," arXiv preprint arXiv:2210.02593, 2022. [Online]. Available: https://arxiv.org/abs/2210.02593

[12] 6. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, Y. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, 2020. [Online]. Available: https://arxiv.org/abs/2005.1140

[13] 7. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” EMNLP Workshops, 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[14] 8. H. Johnson, M. Douze, and H. Jégou, “Billion-Scale Similarity Search with GPUs,” FAISS Documentation (Facebook AI), 2020. [Online]. Available: https://faiss.ai/

[15] 9. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT, 2019. [Online]. Available: https://aclanthology.org/N19-1423