Applied Data Science Department

Multimodal Content Generation using LLM

Project Advisor: Simon Shim

Bashaarat Nawaz Mohammad
Dilip Kumar Kasina
Soumiya Rao Thada
Vaishnavi Samboji
Venkata Kiran Reddy Kotha

Introduction

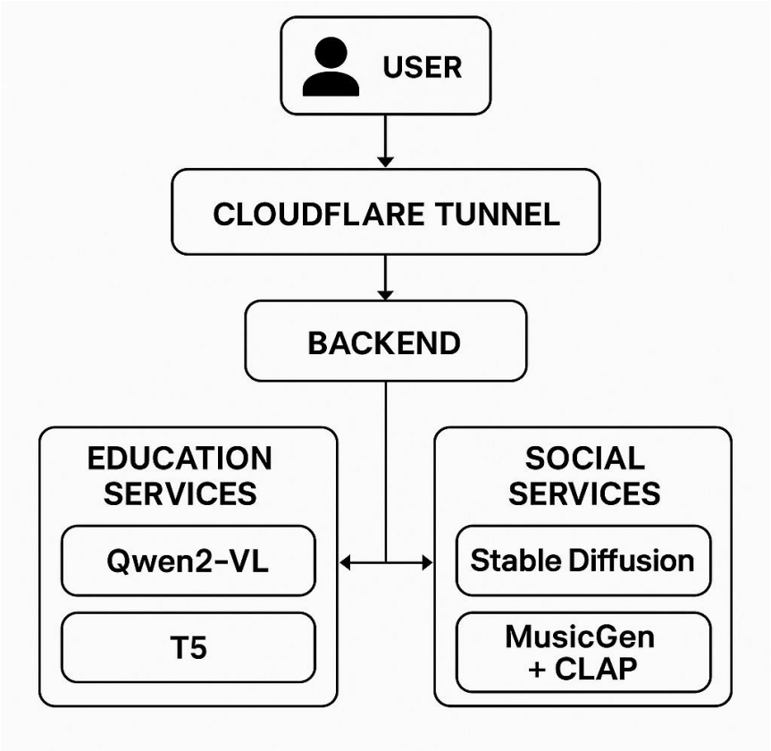
Digital content underpins both social media and learning, but creating it manually is slow and skill-intensive. This project builds a unified Multimodal Content Generator that uses fine-tuned open-source models (Stable Diffusion, MusicGen-small, Qwen VLM, T5-small) to generate social-media images and music, answer multimodal science questions, and summarize research-style text.

ContentForge AI

- Two domains in one system: Social Media (image + music) and Education (QA + summarization).
- Fine-tuned on ScienceQA, arXiv summarization data, and a genre-labeled music dataset.
- Automates key content creation and study tasks, reducing manual effort.
- Exposed through a modular backend and web UI for end-to-end use.

Methodology

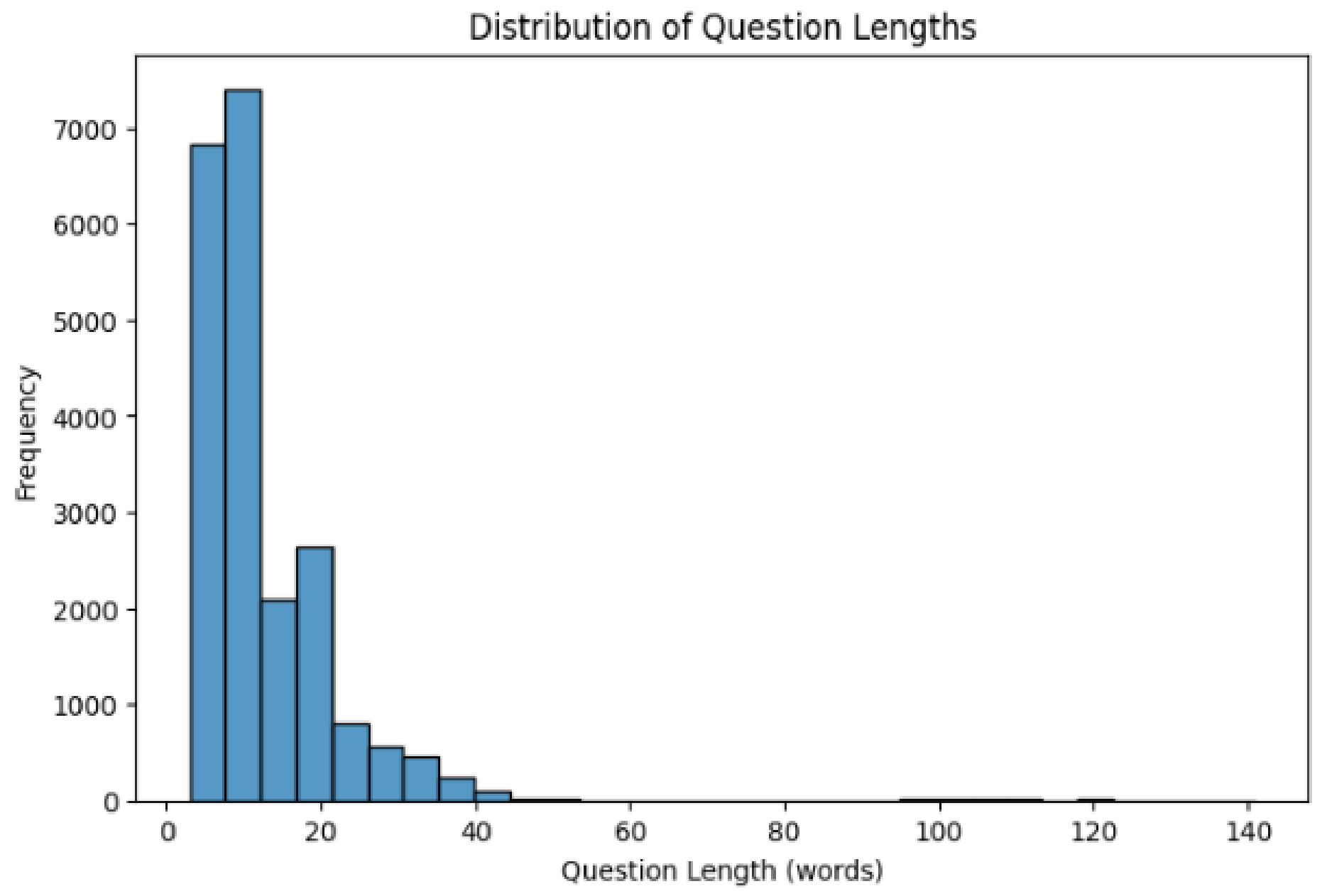
We followed an iterative, experiment-driven workflow that moved from model fine-tuning to system integration and UI. Instead of a strict linear pipeline, we cycled between small training runs, metric-based evaluation, and vertical end-to-end tests that connected data, models, backend, and frontend.



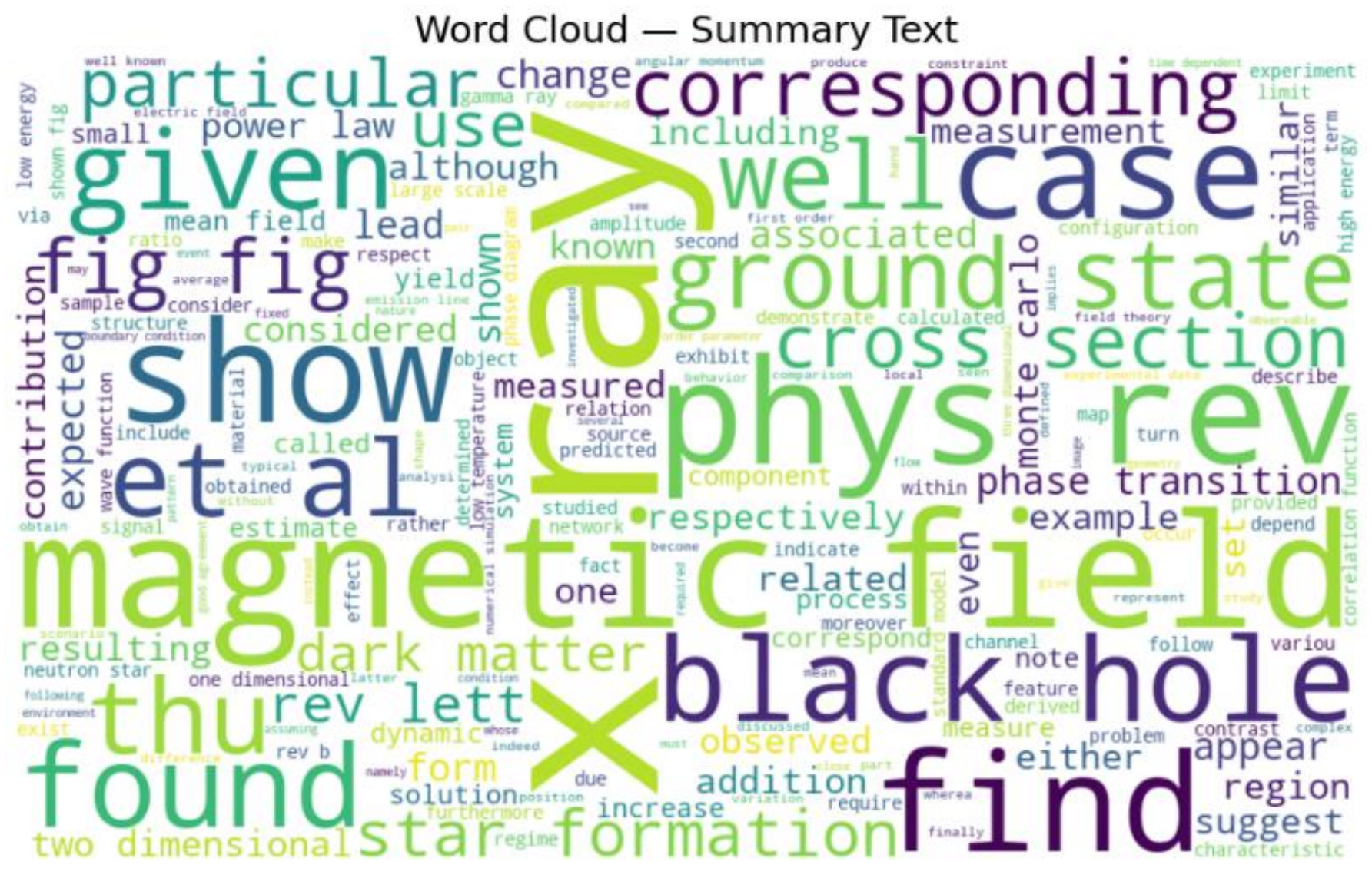
- Built and validated training pipelines for Qwen VLM, T5-small, and MusicGen-small using prepared ScienceQA, arXiv, and music datasets.
- Ran repeated fine-tuning experiments, adjusting learning rates, batch sizes, and selecting configurations based on validation metrics and qualitative outputs.
- Wrapped each model behind a backend service and added a controller to route requests for image generation, music generation, multimodal QA, and summarization.
- Implemented **vertical slices** from UI → backend → model → output, then refined both models and interface based on end-to-end testing.

Analysis and Results

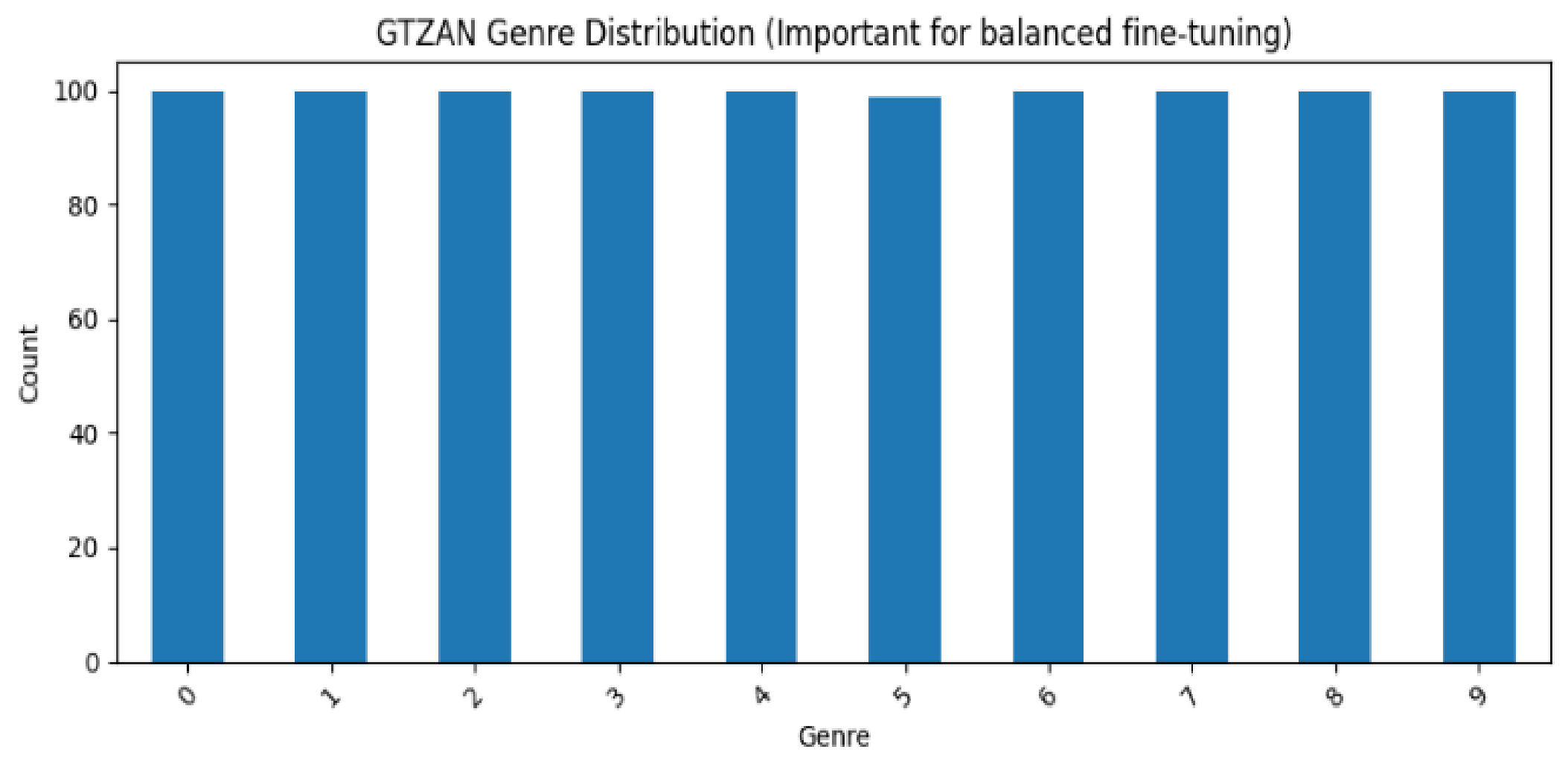
We analysed the three main datasets that drive the system before fine-tuning any models. For **ScienceQA**, the report examines how questions are distributed across subjects (natural, social, and language science) and how many items include text context, image context, or both. Length statistics for questions, lectures, and explanations were computed to understand typical sequence sizes. These analyses informed tokenizer settings and maximum input lengths for Qwen VLM, ensuring that multimodal questions with diagrams and explanatory text could be processed without excessive truncation.



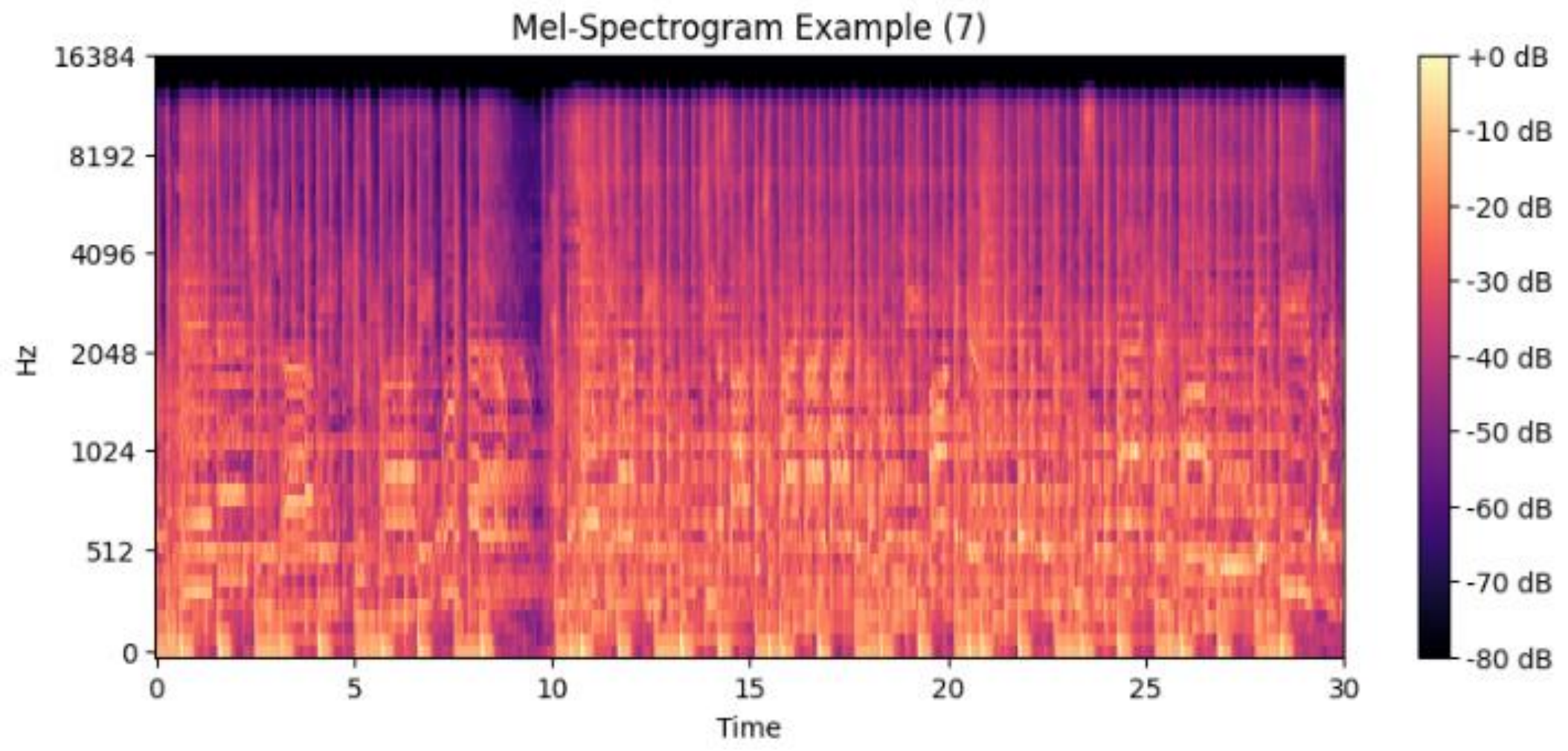
For the **arXiv summarization corpus**, the report compares article and summary lengths and looks at the ratio between them. Histograms show that articles are substantially longer than their corresponding abstracts, confirming that the task requires strong compression while preserving key information. Token-length plots for inputs and outputs guided choices of encoder and decoder limits for T5-small, balancing coverage of long research sections against GPU memory constraints.



For the **music dataset**, based on GTZAN-style clips, the report summarizes the number of examples per genre and explores basic audio characteristics such as duration and time–frequency structure. Genre distribution plots and mel-spectrogram visualizations provide intuition about the diversity of training material that MusicGen-small must learn to reproduce and adapt to text prompts.



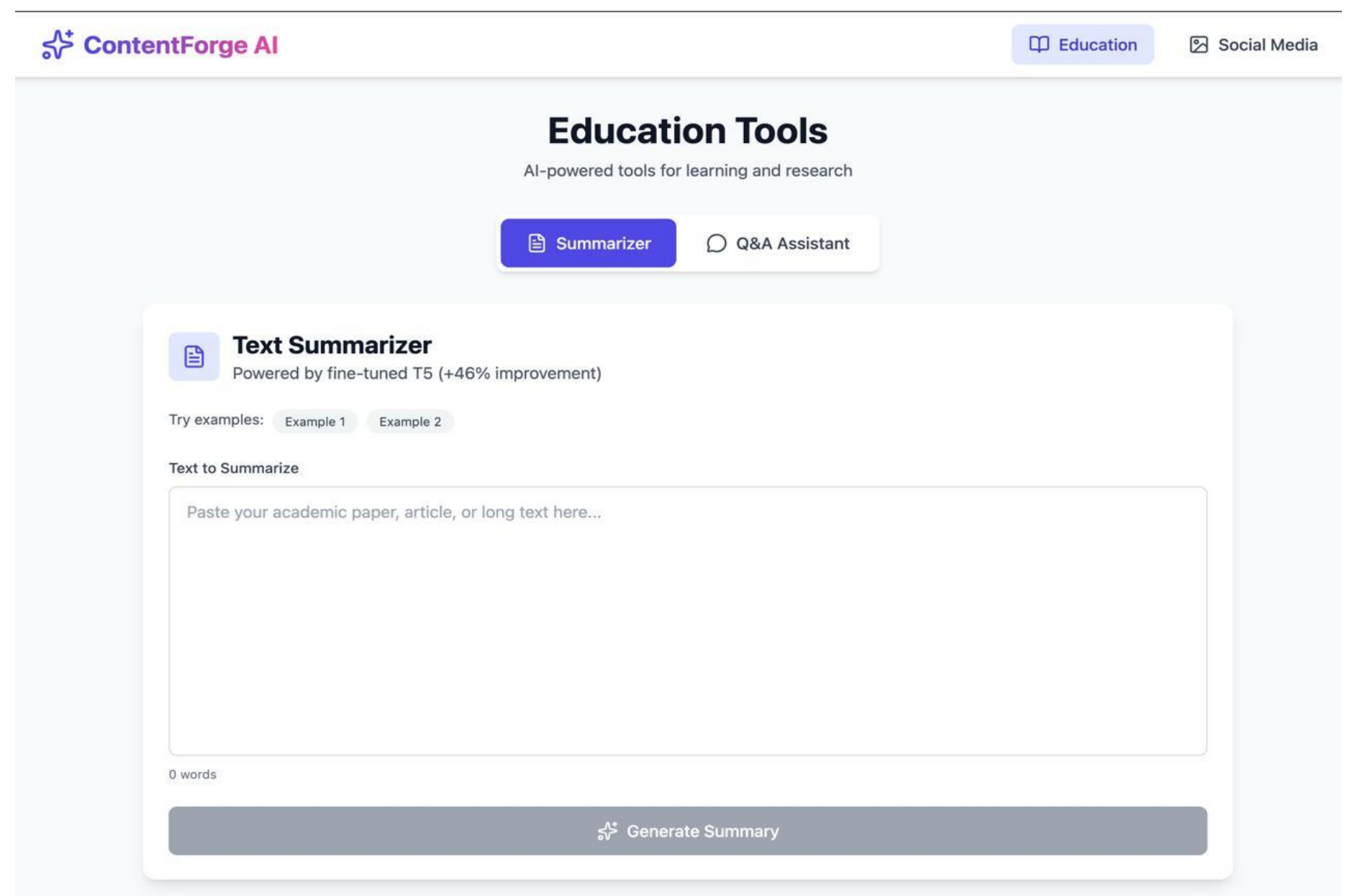
Model evaluation compares base and fine-tuned versions of each component. Fine-tuned **T5-small** on the arXiv corpus improves all reported summarization metrics—ROUGE variants, BLEU, METEOR, BERTScore, FactCC, and perplexity—relative to the base model, and qualitative examples show that summaries become more focused and technically appropriate. For **MusicGen-small**, CLAP-guided fine-tuning yields better FAD_CLAP scores and higher CLAP text–audio similarity, indicating that generated clips align more closely with requested moods and genres. Fine-tuned **Qwen VLM** achieves higher answer accuracy on the ScienceQA evaluation split, especially for questions that combine text and diagrams, with qualitative examples illustrating more grounded reasoning over visual content. **Stable Diffusion** outputs, evaluated qualitatively in the report, produce relevant social-media thumbnails and educational illustrations for a range of prompts, demonstrating that the image component can support both domains.



Fine-tuning Qwen2-VL on ScienceQA raised accuracy from about 65% to ≈67–68%, with bigger gains on image+text questions. For T5-small on arXiv, ROUGE-1 improved 0.22 → ≈0.28, ROUGE-2 0.056 → ≈0.083, BLEU 0.0023 → 0.0096, and perplexity 7.63 → 1.52. CLAP-guided MusicGen-small reduced FAD_CLAP ≈0.97 → ≈0.91 and achieved the strongest text–audio alignment among the tested music models.

Taken together, these analyses and results show that carefully chosen datasets plus targeted fine-tuning lead to measurable gains over base models, and that the integrated system can handle realistic multimodal inputs for social media content creation and science learning.

User Interface



Education Tools

AI-powered tools for learning and research

Text Summarizer

Powered by fine-tuned T5 (+46% improvement)

Try examples:

Text to Summarize

Paste your academic paper, article, or long text here...

0 words

Summary/Conclusions

This project builds a **Multimodal Content Generator** that unifies Stable Diffusion (images), MusicGen-small (music), Qwen VLM (multimodal QA), and T5-small (scientific summarization) behind a modular backend and simple web UI.

Fine-tuned models improve summarization quality, text–audio alignment, and QA accuracy over base versions, while remaining limits in compute, music data size, and occasional errors suggest clear directions for future improvement.

Key References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of CVPR 2022.

[2] Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). MusicGen: Simple and Controllable Music Generation. arXiv preprint arXiv:2306.05284.

[3] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., & Zhou, J. (2023). Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966.

[4] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1–67.

Acknowledgements

We would like to sincerely thank Dr. Simon Shim for his guidance and support throughout this project. His feedback helped us refine our problem formulation, choose appropriate models and datasets, and interpret our experimental results. We are also grateful to the course staff and our classmates for their discussions and suggestions, which helped us debug issues, improve our system design, and keep the project on track.