Reasoned Tokenization Across Modalities: Processing ASCII Art and MusicXML Through Fundamental NLP Techniques

Kevin Hopkins[1], **Adam Lehavi**[1], **Andy Chen**[1], **Jeffrey Yeh**[1], **Chengxi Xu**[1], **Peter Xu**[1]
[1]University of Southern California
{kevinhop,alehavi,chenandy,jtyeh,chengxix,xuzihao}@usc.edu

Abstract

We present a comprehensive study investigating how unimodal language models can interpret non-traditional text representations—including ASCII art and MusicXML—when each modality is strictly processed as token sequences. We focus on LLaMA3.2(1B parameters) augmented with an Entropix sampling strategy, testing its ability to classify ASCII-based visual structures and symbolic music content without resorting to multimodal architectures. Our approach includes brightness-based encoding for ASCII art, few-shot prompting, and text-only processing of MusicXML data. Experimental results on ASCII classification highlight consistent gains from brightness encoding, revealing that token-level representations of spatial patterns can boost performance. In contrast, MusicXML classification results are more limited, reflecting the inherent challenges of tokenizing complex musical attributes. Taken together, these findings show that fundamental NLP mechanisms—tokenization, embedding, and next-token prediction—retain surprising versatility across diverse text-like inputs, although domain-specific encodings and targeted fine-tuning remain critical for robust performance in non-traditional textual domains.

1 Introduction

Recent advances in large language models (LLMs) have driven major improvements in conventional natural language processing (NLP) tasks. Yet when it comes to non-textual data formats like images, audio, or musical scores, standard practice often gravitates toward multimodal architectures or specialized encoders that convert these modalities into embeddings more palatable for language models. In contrast, our work explores a straightforward but provocative question: *Can unimodal language models (LMs) interpret structured, non-traditional text such as ASCII art and MusicXML purely by treating them as token sequences?*

ASCII art naturally embeds spatial information through arrangements of printable characters, while MusicXML provides a hierarchical, text-based representation of symbolic music. Traditional methods might convert ASCII grids into pixel images or parse MusicXML with a specialized music-processing pipeline, introducing significant architecture overhead. Here, we pursue an alternative: directly adapting fundamental NLP techniques—tokenization, embeddings, next-token prediction, and fine-tuning—to classify and reason about ASCII and MusicXML purely as textual tokens.

Our main contributions are:

- **Brightness-based ASCII encoding.** We present a novel encoding pipeline that translates ASCII layouts into brightness values. This strategy boosts classification performance over naive baselines, reaffirming that direct text processing can be surprisingly effective for visual patterns.
- **MusicXML genre classification.** We outline a text-only pipeline for symbolic music classification, achieving modest gains over vanilla LLaMA but uncovering challenges unique to highly structured musical data.
- **Reasoning enhancements via Entropix sampling.** Fine-tuning smaller models with Entropix (Phogat et al., 2024) yields partial improvements in classification consistency, showcasing how targeted training bolsters reasoning.
- **Cross-modal analysis.** We find that while token-based approaches can handle ASCII's visual structure, symbolic music demands a more domain-sensitive strategy. Our findings underscore both the promise and the limitations of applying core NLP mechanisms to non-traditional textual modalities.

In short, our results suggest that unimodal LLMs can indeed parse and reason about alternative text representations without the overhead of multimodal architectures, but domain-targeted encodings and careful fine-tuning remain crucial. The remainder of this paper details our entire pipeline—from dataset curation and token encoding to final classification metrics—and closes with an analysis of cross-modal reasoning in unimodal LLMs.

2 Related Work

ASCII Art in NLP. Recent studies have begun to highlight the unique challenges ASCII art poses for language models, particularly in recognizing spatial arrangements encoded within text. Jiang et al. (2024) showed that ASCII art can exploit gaps in alignment by functioning as an adversarial "jailbreak" mechanism. Their findings suggest that LLMs struggle not only to *interpret* ASCII art accurately but also to maintain alignment policies in the face of non-standard textual patterns. Our work complements this perspective by focusing on *positive* classification and reasoning performance—i.e., systematically parsing ASCII layouts—rather than adversarial prompting.

Fine-Tuning Smaller Language Models. Techniques for boosting the reasoning capabilities of modestly sized LLMs form a key component of our approach. Phogat et al. (2024) demonstrated that fine-tuning smaller models on targeted question-answering datasets yields improvements in logical inference and structured reasoning. Drawing from these insights, we combine standard reasoning benchmarks (e.g., LAMBADA, LogiQA) with an entropy-based "Entropix" sampling strategy to enhance classification consistency in our 1B-parameter LLaMA model. This approach specifically targets the ability to interpret spatial patterns in ASCII art and symbolic structures in MusicXML.

Symbolic Music Processing. Symbolic music representations, such as MusicXML, have garnered interest in both NLP and music information retrieval contexts. Long et al. (2024) introduced the *PDMX* dataset, a large-scale collection of public domain MusicXML files, which we leverage for genre classification. While prior methods commonly rely on specialized symbolic music encoders or convert the data to MIDI, our study shows that straightforward text-based processing—augmented with only minor modifications for domain-specific structure—can still yield informative results.

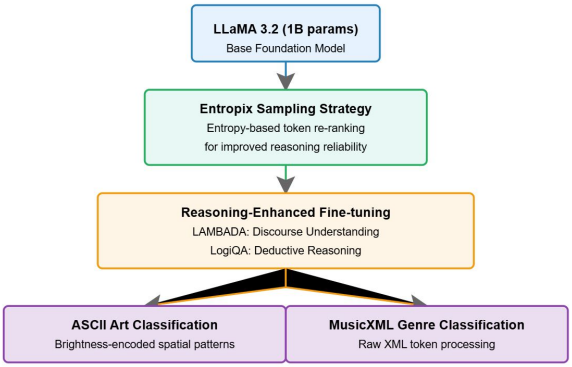

Figure 1: High-level architecture of our approach. LLaMA 3.2 (1B params) serves as the base model, supplemented by an Entropix sampling strategy for improved token re-ranking. We then apply reasoning-enhanced fine-tuning on LAMBADA (discourse understanding) and LogiQA (deductive reasoning). Finally, we test the model on ASCII art (with brightness encoding) and raw MusicXML classification.

Reasoning Datasets and Benchmarks. Finally, a variety of established reasoning datasets guide the development of our methods. We incorporate LAMBADA (Paperno et al., 2016) to improve broad discourse understanding and LogiQA (Liu et al., 2020) to bolster deductive reasoning, as well as ARC-Easy. Rather than applying these resources purely to traditional language tasks, we use them to enhance cross-modal reasoning on ASCII art and symbolic music, underscoring the broader applicability of standard NLP benchmarks to unconventional tokenized domains.

3 Project Overview

Our project explores how unimodal language models can parse non-traditional text modalities by leveraging straightforward tokenization and fine-tuning methods. We build on a 1B-parameter LLaMA 3.2 base model and introduce an Entropix sampling strategy as well as additional reasoning-enhanced fine-tuning steps.

In this pipeline, the base model benefits from an entropy-aware sampling mechanism (Entropix) to improve reliability of structured reasoning, followed by domain-specific tuning on ASCII art and MusicXML tasks.

Our project investigates whether fundamental NLP constructs—tokenization, embeddings, next-token prediction, and reasoning-oriented fine-

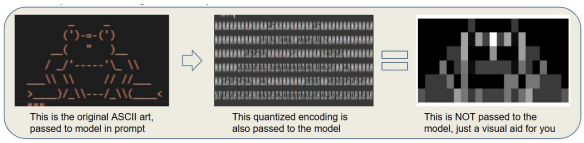

Figure 2: Visualization of our brightness-based encoding pipeline for ASCII art. (Left) The raw ASCII art. (Center) Quantized brightness values represented as numeric tokens. (Right) A conceptual grayscale matrix for illustration. Only the numeric tokens are passed into the language model.

tuning—can directly parse ASCII art and MusicXML without specialized or multimodal architectures. Three core principles guide this exploration:

Text-only Approach. Rather than converting ASCII to pixel images or MusicXML to specialized symbolic formats, we preserve *all* data in its raw text form. This choice leverages standard NLP pipelines and avoids introducing additional modalities.

Lightweight Model. We rely on a 1B-parameter LLaMA 3.2 model, balancing computational feasibility with representational capacity. By testing on a modestly sized LM, we highlight the scalability potential for smaller architectures when applied to non-traditional text.

Reasoning Enhancements. We enhance structured reasoning via Entropix sampling (Phogat et al., 2024) and complementary fine-tuning steps on LAMBADA (Paperno et al., 2016) and LogiQA (Liu et al., 2020). These techniques reinforce pattern recognition and logical inference in both ASCII art and symbolic music contexts.

4 Technical Approach

We explore the viability of applying standard NLP principles to non-traditional text data—specifically ASCII art and MusicXML—using a LLaMA 3.2 1B architecture. Below, we describe our specialized encoding pipelines for each modality and the fine-tuning strategies used to enhance the model's reasoning capabilities.

4.1 ASCII Art Encoding

ASCII art is text-based yet encodes visual information through spatially arranged characters. To leverage standard NLP workflows, we designed the following two-step process:

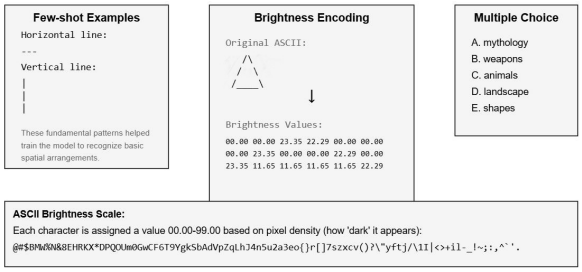

Figure 3: Our overall ASCII art prompt structure. (Left) A few-shot demonstration of fundamental patterns, such as horizontal and vertical lines, used to teach basic spatial arrangements. (Center) A sample brightness encoding of simple ASCII artwork, where each character is assigned a value (00.00–99.00) based on pixel density. (Right) Multiple-choice classification labels for final output. The ASCII Brightness Scale at bottom shows how different characters map to denser or lighter numeric values.

Grid Standardization. We scraped 5,225 ASCII samples from `www.asciiart.eu`. Each piece was standardized by trimming extra whitespace, normalizing line lengths, and aligning characters into a consistent rectangular grid. This yields a uniform "canvas" that preserves the ASCII art's spatial structure.

Brightness Encoding. To reflect visual patterns in purely textual form, each character is mapped to a brightness score on a 0–100 scale. Denser characters pixel-wise (like '@' or '8') map to higher brightness values, while spaces map to zero. The resulting brightness matrix—essentially a text-based intensity grid—is serialized into tokens and fed directly into our model.

Figure 2 shows a conceptual illustration of this pipeline, while Figure 3 provides a more detailed look at our ASCII art prompt structure, including few-shot examples of simple lines, the brightness encoding transformation, and the multiple-choice classification setup.

Prompt Structure. In our classification experiments, we embed both the raw ASCII snippet and its brightness-encoded matrix into a single prompt, accompanied by few-shot examples of basic patterns (e.g., horizontal/vertical lines) as shown in Figure 3. The prompt concludes with multiple-choice labels (e.g., `A. animals, B. shapes, etc.`), from which the model selects the best match.

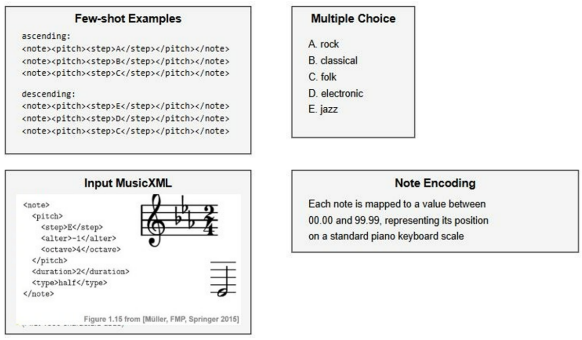

Figure 4: Our MusicXML processing pipeline and prompt structure. (Top-left) Few-shot examples for ascending and descending note sequences. (Top-right) Multiple-choice classification labels (e.g., `A. rock, B. classical, etc.`). (Bottom-left) Snippet of input MusicXML along with a visual snippet (Müller , 2015). (Bottom-right) Our note encoding scheme, mapping each pitch to a numeric token (00.00–99.99) based on piano keyboard scale. All components are passed to the model as a single prompt sequence.

4.2 MusicXML Processing

MusicXML represents symbolic music via XML-style tags encoding pitches, durations, and instrumentation. We rely on the PDMX dataset (Long et al., 2024) for single-track compositions:

Data Filtering and Balancing. From the original 35,000-file corpus, we isolate single-track pieces with clearly labeled genres. For consistency across genres, we sample approximately 30 compositions per class, leading to a balanced subset for classification tasks.

Direct Text Input. Each MusicXML file is fed as raw text, preserving nested tags such as `<measure>` and `<pitch>`. We deliberately avoid specialized music embeddings or staff-based interpretations, instead treating MusicXML content purely as token sequences.

Prompt Structure. As shown in Figure 4, we embed few-shot examples (e.g., ascending/descending pitch sequences) plus the raw MusicXML snippet into a single prompt. Each note is mapped to a numeric value corresponding to its pitch, and the model is then prompted to select the correct genre label from multiple choices.

4.3 Model Architecture and Fine-Tuning

Base Model. All experiments are conducted using LLaMA 3.2 with 1B parameters, hosted under *meta-llama/Llama-3.2-1B-Instruct*. This relatively small architecture allows us to assess efficiency and scalability for non-traditional token processing.

Entropix Sampling. We incorporate an entropy-based sampling strategy proposed by Phogat et al. (2024), which re-ranks candidate tokens during generation to mitigate hallucinations. This mechanism aims to improve structured reasoning by promoting more reliable token choices in uncertain contexts.

Reasoning Datasets. The model is further fine-tuned on LAMBADA (Paperno et al., 2016) for broad discourse reasoning and LogiQA (Liu et al., 2020) for logical inference, as well as ARC-Easy. These tasks reinforce the model's ability to interpret spatial and symbolic patterns, bridging the gap between standard language data and ASCII or MusicXML token sequences.

5 Datasets and Experiments

5.1 ASCII Art Dataset

Data Curation. We scraped 5,225 ASCII art samples from `www.asciiart.eu`, each labeled with a top-level category (`animals`, `vehicles`, `plants`, etc.). After removing classes with fewer than 30 samples and merging certain subcategories (e.g., `dogs`, `cats`, and `birds` into `animals`), we split the dataset into an 80% training set and 20% test set.

Encoding Approaches. We investigate two tokenization pipelines:

1. **Baseline:** Input the raw ASCII art directly with minimal token preprocessing.

2. **Brightness Grid:** Map each character to a numerical brightness score (0–100), effectively emulating image intensity values while preserving a text-only modality.

These two representations are passed to the model in the prompt, allowing direct comparison between naive and domain-targeted encodings.

Classification Setup. Each ASCII art piece is classified via a multiple-choice prompt, which includes the ASCII snippet (and/or brightness matrix) along with a set of possible categories. The model generates a final label token, which is then matched against the ground truth.

5.2 MusicXML Genre Classification

Data Curation. We select single-track MusicXML files from the *PDMX* dataset (Long et al., 2024), focusing on pieces with known genres (e.g., folk, jazz, classical, pop). Approximately 30 files per genre are sampled, reserving 20% for testing to maintain a balanced evaluation.

Prompt Design. To accommodate token length limitations, we truncate each MusicXML input to 1000 characters. Multiple-choice genre labels are appended, and the model must identify which label best matches the musical content. This direct text-based pipeline avoids specialized symbolic music encoders.

Comparison to Vanilla LLaMA. We compare our fine-tuned, Entropix-augmented LLaMA model against a vanilla LLaMA 3.2 1B baseline. This head-to-head evaluation isolates the impact of reasoning-oriented fine-tuning and text-only tokenization on structured music classification tasks.

Direct Music Fine-Tuning We also finetuned separate LLaMA3.21B models exclusively on each modality dataset, using held-out examples from the same domain for testing. This approach shows how well a model can learn domain-specific patterns when trained directly on its target modality. However, it doesn't demonstrate broader cross-modal generalization, since the model is specialized for symbolic music rather than multiple data types.

5.3 Evaluation Metrics

Accuracy and F1. We report both metrics for straightforward assessment of classification performance in ASCII art and MusicXML tasks.

Hierarchical Coverage and Weighted Top-k Accuracy. In the ASCII classification setting, we introduce hierarchical metrics to capture partial credit for near-miss predictions (e.g., predicting a closely related category within the same parent node). This framework reflects the subtlety of class boundaries in ASCII art and emphasizes cross-category relationships.

6 Results and Analysis

6.1 ASCII Art Classification Results

Performance Overview. Our brightness-encoded pipeline consistently outperforms both vanilla LLaMA and a raw ASCII-only pipeline. Pilot experiments show notable gains in standard

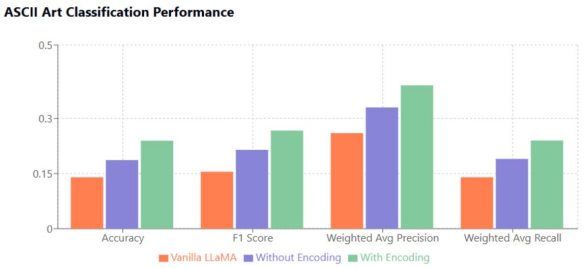

Figure 5: ASCII art classification results comparing Vanilla LLaMA (orange), LLaMA without brightness encoding (blue), and LLaMA with brightness encoding (green). The chart shows that brightness encoding yields consistently higher performance across Accuracy, F1 Score, Weighted Avg Precision, and Weighted Avg Recall.

metrics: the best brightness-encoded model achieves **23.94%** accuracy and an F1 score of **26.70%**, compared to **18.66%** accuracy and **21.45%** F1 without encoding. The vanilla LLaMA baseline lags at **14%** accuracy and **15.5%** F1. This improvement appears to stem from the model's ability to leverage numeric brightness tokens, which provide a more structured representation of spatial layout.

Reasoning Demonstration. Qualitative analysis reveals that the fine-tuned model reliably recognizes fundamental ASCII patterns (e.g., horizontal vs. vertical arrangements). We hypothesize that reasoning-oriented fine-tuning on datasets like LAMBADA helps the model derive contextual meaning from token sequences—even when those tokens represent spatial information rather than natural language.

6.2 MusicXML Genre Classification Results

Modest Gains. Compared to ASCII art classification, the model's improvements on MusicXML are more modest. Despite fine-tuning, our best configuration *with encoding* achieves **22.92% accuracy** and an F1 score of **22.57%** on a multi-class genre task—marginally above the **22.22% accuracy** (F1: **21.99%**) of the vanilla LLaMA baseline. Meanwhile, our *non-encoded* fine-tuned model trails at **19.54% accuracy** (F1: **19.34%**).

Analysis of Shortcomings. MusicXML data encodes intricate musical concepts (pitch, rhythm, instrumentation) through verbose XML tags. Without a domain-specific approach akin to brightness encoding, many critical structural and musical cues remain untapped. Truncating MusicXML files to

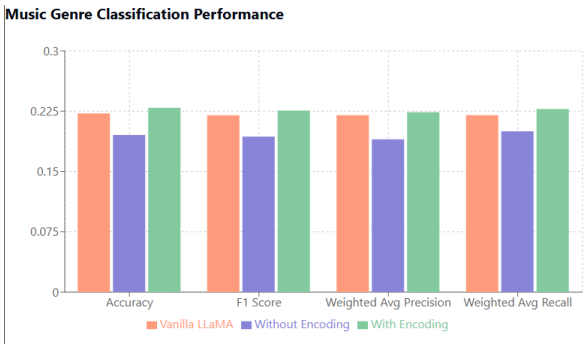

Figure 6: Music genre classification performance comparing our fine-tuned model with (green) and without encoding (purple), along with Vanilla LLaMA (orange). We observe modest improvements in Weighted Avg Precision (22.37%) and Weighted Avg Recall (22.77%) with encoding, but overall gains remain limited compared to ASCII art classification.

1,000 characters further compounds this issue, as key information can be lost.

Negative Result Interpretation. The gap between ASCII and MusicXML outcomes underscores that not all structured text benefits equally from the same NLP adaptations. While brightness encoding significantly boosted ASCII classification, no corresponding domain-specific encoding has been applied to MusicXML. Our findings suggest that purely textual handling of symbolic music without specialized embeddings or extended context may not suffice for higher accuracy.

Direct Fine-Tuning Observation. We also evaluated the model performance on a task when fine-tuned directly on the dataset of that modality. This domain-specific model achieved **0.2375** accuracy and an **F1 score of 0.2378** on the MusicXML classification task and an accuracy of **0.200** on the ASCII task. This domain-specific model performs well on its associated modality with music classification. Instead of indicating cross-modal generalization, this illustrates that training on a modality provides a performance boost in that modality.

6.3 Cross-Modal Observations

Collectively, our experiments confirm that straightforward NLP pipelines excel when ASCII's spatial patterns are translated into brightness tokens, but are less effective on symbolic musical notation without tailored token engineering. These contrasting results highlight that while unimodal LLMs can adapt to some non-traditional text domains with relatively simple modifications, domain-targeted encodings and careful fine-tuning remain vital for robust cross-modal generalization.

7 Discussion and Future Work

Our findings illustrate both the potential and the limitations of applying unimodal language models to highly structured, non-traditional textual domains. While straightforward encoding strategies succeeded with ASCII art, genre classification on MusicXML lagged behind. Below, we analyze the core factors behind these discrepancies and outline steps for advancing text-only approaches in more complex modalities.

7.1 Disparate Results Across Modalities

One striking outcome of our experiments is the disparity between strong performance gains in ASCII art classification and only marginal improvements in MusicXML. ASCII inherently exists in a text-based form that closely mirrors how LLMs process sequential data; spatial patterns (e.g., horizontal or vertical lines) remain visually meaningful even when tokenized. In contrast, MusicXML encodes intricate musical concepts (e.g., pitch, duration, instrument) through verbose XML tags, making it considerably less direct to interpret as plain text.

For instance, ASCII art brightness encoding aligns naturally with the notion of image intensity, yielding numeric tokens the model can readily ingest and correlate with spatial structure. However, an analogous approach for music would likely require designing specialized, multi-dimensional encodings (e.g., separate embeddings for pitch, duration, and measure boundaries) rather than relying on raw XML tags. As a result, representing MusicXML purely as plain text forces the model to parse a hierarchical structure that does not map cleanly onto sequential token embeddings. Consequently, the unimodal LM struggles to capture crucial musical relationships, resulting in smaller gains.

These divergent outcomes suggest that not all "structured text" is equally suitable for direct tokenization. Modalities like ASCII art, which already manifest as coherent character grids, can be encoded in ways that play to the strengths of an LLM's token-based processing. Other modalities, including symbolic music, video transcripts, or even certain scientific data formats, may require more domain-targeted transformations—or possibly larger context windows—to realize the same

benefits.

7.2 Classification Results

As shown in Figure 5, brightness encoding significantly boosts performance across all metrics compared to both vanilla LLaMA and the unencoded ASCII baseline. We observe a notable jump in Weighted Avg Precision (approximately 0.28 vs. 0.24), reflecting the model's improved ability to distinguish between categories.

Meanwhile, Figure 6 illustrates that our fine-tuned model yields only marginal gains over vanilla LLaMA on MusicXML in terms of F1 Score and Weighted Avg Recall. The lack of clear improvements suggests that symbolic musical data—unlike ASCII art—demands more sophisticated encodings and possibly deeper fine-tuning for the model to discern relevant patterns.

Structured Encodings Drive Performance. The ASCII brightness-encoding pipeline exemplifies how a domain-targeted transformation can significantly boost classification. Analogous approaches may be directly transferable to MusicXML, where encoding pitch classes, note durations, or hierarchical measure boundaries could better capture musical structure. Our negative results on raw MusicXML underscore that off-the-shelf tokenization often fails to capture domain-specific nuances.

Scaling Reasoning to Specialized Domains. While Entropix sampling modestly improved classification consistency, those gains were limited in MusicXML tasks. This shortfall suggests that broader coverage of music-centric reasoning examples—akin to how ASCII tasks were supplemented with explicit pattern prompts—could be necessary to fully leverage reasoning-oriented fine-tuning.

Lightweight LLM Deployments. Our results show that a 1B-parameter LLaMA model can effectively handle some forms of structured text with proper domain encodings. This finding highlights the viability of compact, resource-friendly solutions for specialized tasks like ASCII classification. However, the MusicXML setting exposes the inherent limits of smaller architectures without more targeted preprocessing.

7.3 Future Work

Future research could explore music-specific encoding strategies akin to the brightness-based pipeline demonstrated for ASCII art. For instance, designing tokens that capture pitch classes, measure boundaries, and note relationships might yield clearer structural representations of MusicXML data. Another direction involves comparing a reasoning-finetuned model (as described in this work) against a model trained on conventional language tasks, examining whether incorporating non-traditional text modalities impacts performance on standard benchmarks. Researchers might also investigate different fine-tuning setups—from few-shot prompting to hierarchical tokenization—to better integrate complex, domain-specific structures into purely textual embeddings.

8 Conclusion

We have demonstrated that unimodal language models can parse and classify both ASCII art and MusicXML purely as textual input, provided domain-specific encodings and careful fine-tuning strategies are applied. In particular, brightness-based encoding for ASCII art yielded consistent gains in classification accuracy, highlighting how small architectural modifications can leverage LLMs' next-token prediction abilities for spatially structured text. MusicXML classification, however, remains challenging when treated as raw tokens: without a parallel encoding scheme or larger context windows, the model struggled to fully capture musical structure. Together, these findings reinforce the broader insight that while fundamental NLP methods—tokenization, embeddings, and fine-tuning—are highly adaptable, *specialized encodings and context-aware strategies* remain critical for success in structured domains beyond natural language.

Acknowledgments

We thank the University of Southern California for providing organizational resources, as well as the maintainers of `www.asciiart.eu` and the PDMX dataset (Long et al., 2024) for making valuable public data available.

References

Jiang, F., Xu, Z., Niu, L., Xiang, Z., Ramasubramanian, B., Li, B., and Poovendran, R. (2024). ArtPrompt: ASCII Art-based Jailbreak Attacks against Aligned LLMs. *arXiv:2402.11753[cs.CL]*.

Liu, J., Cui, L., Liu, H., Huang, D., Wang, Y., and Zhang, Y. (2020). LogiQA: A Challenge Dataset

for Machine Reading Comprehension with Logical Reasoning. *arXiv:2007.08124[cs.CL]*.

Long, P., Novack, Z., Berg-Kirkpatrick, T., and McAuley, J. (2024). PDMX: A Large-Scale Public Domain MusicXML Dataset for Symbolic Music Processing. *arXiv:2409.10831[cs.SD]*.

Müller (2015). Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. https://doi.org/10.1007/978-3-319-21945-5.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. (2016). The LAMBADA dataset: Word prediction requiring a broad discourse context. *arXiv:1606.06031[cs.CL]*.

Phogat, K. S., Puranam, S. A., Dasaratha, S., Harsha, C., and Ramakrishna, S. (2024). Fine-tuning Smaller Language Models for Question Answering over Financial Documents. *arXiv:2408.12337[cs.CL]*.