NTU MIR 2025 - Homework 3

Student: 邱冠銘

Student ID: R14921046

Source Code and Results: https://drive.google.com/drive/folders/1v4Iu2vFVS8-GRvOC9mBUO7goI4JZzTmb?usp=sharing

Task 1: Unconditional Generation

Overview

Train a transformer-based model from scratch to generate 32-bar symbolic music using autoregressive generation with different model architectures and tokenization schemes.

Evaluation Metrics

1. **H4** - Pitch-Class Histogram Entropy (4 bars): measures erraticity of pitch usage
2. **GS** - Grooving Pattern Similarity: measures consistency of rhythm across the piece

Token Representation

REMI

- Bar, Position, Pitch, Duration, Velocity
- Explicit bar markers
- Clear temporal structure

REMI+

- Extended REMI with chord tokens
- Additional musical context
- Better harmonic representation

MIDILike

- Note-On, Note-Off, Time-Shift
- More compact representation
- Event-based encoding

Model Architecture

GPT-2 (124M parameters)

- 12 layers, 12 heads
- 768 hidden dimensions
- Standard transformer decoder

BabyLM-GPT2 (100M parameters)

- Pre-trained on developmentally plausible corpus
- 12 layers, 12 heads
- Better initialization for language modeling

Common Settings

- **Max sequence length:** 1024 tokens
- **Vocabulary size:** ~400-500 tokens (varies by tokenizer)

Training & Inference Configuration

Training

Parameter	Value
Epochs	100
Batch Size	8
Learning Rate	0.0004
Optimizer	AdamW
Weight Decay	1e-5
LR Scheduler	One Cycle
Dataset	Pop1K7 (1747 files)

Inference

Parameter	Value
Top-k	5
Temperature	1.2
Repetition Penalty	1.2
Target Bars	32
Max Length	1024

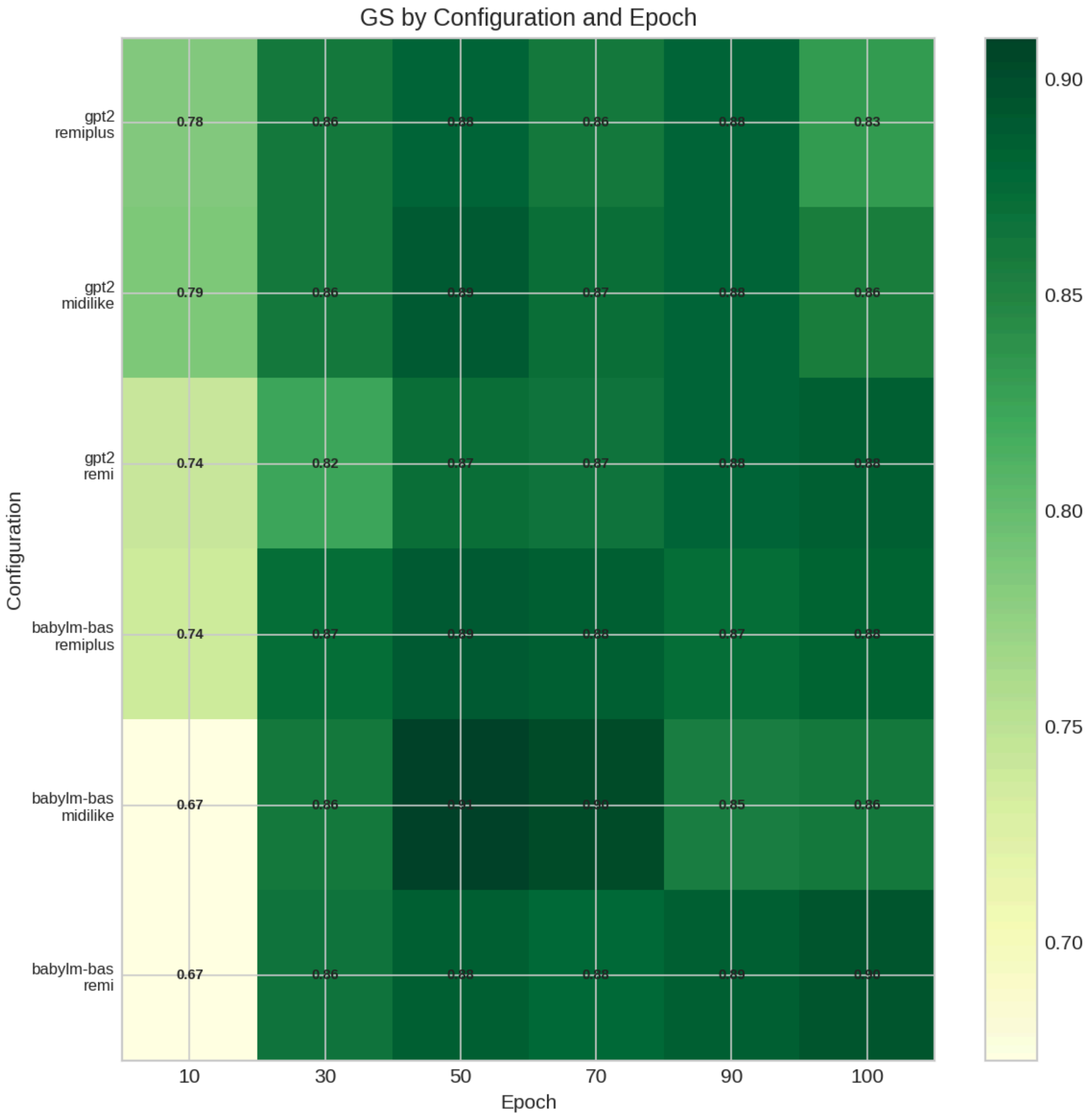
All Experiment Results (Epochs 10-100)

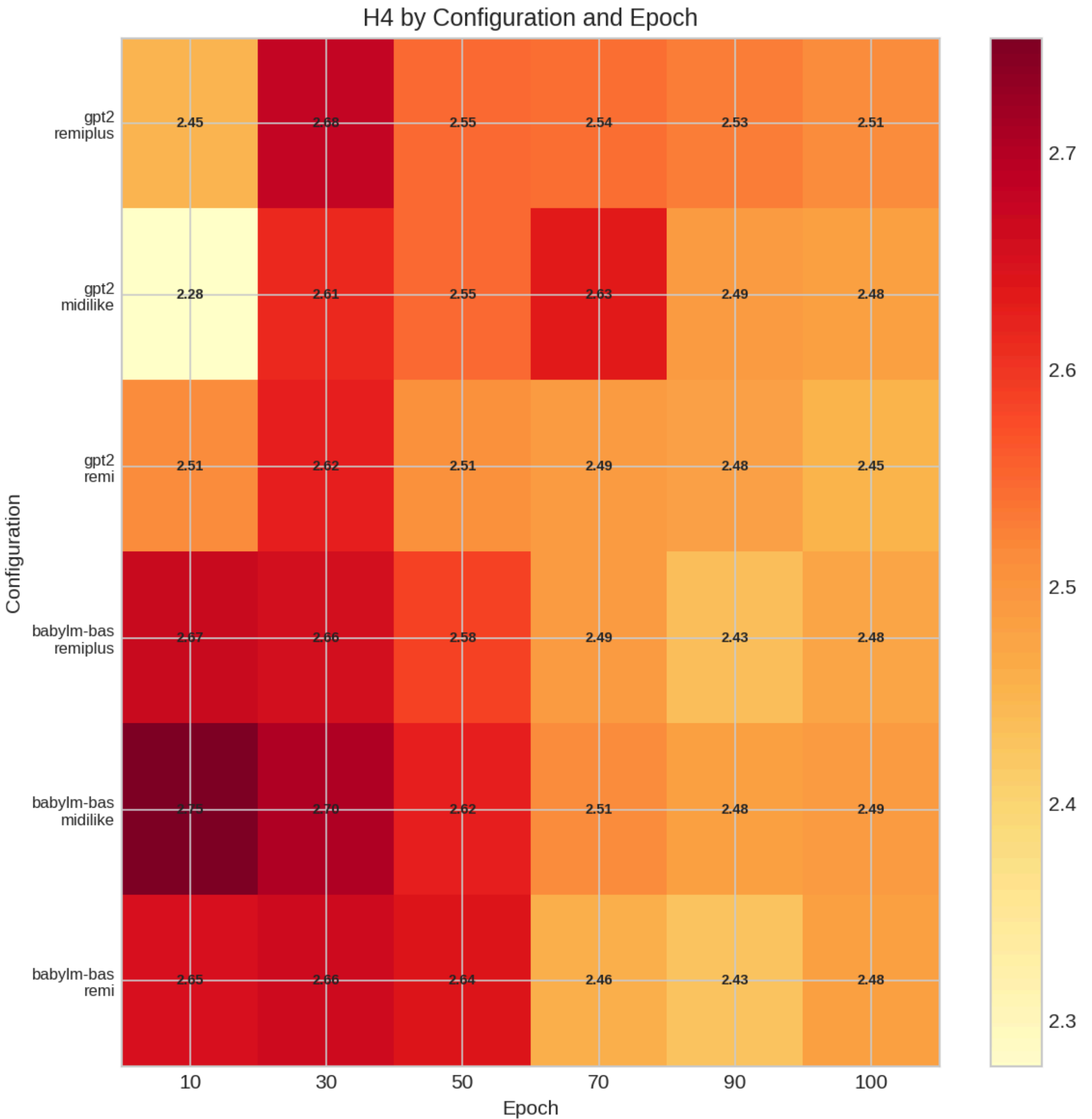
Model	Representation	Event	Loss	Epoch	Top-k	Temp	H1	H4	GS	SI_short	SI_mid	SI_long
gpt2	remiplus	note-on/off	CE	10	5	1.2	1.634	2.446	0.783	-	-	-
gpt2	remiplus	note-on/off	CE	30	5	1.2	1.909	2.681	0.859	-	-	-
gpt2	remiplus	note-on/off	CE	50	5	1.2	1.721	2.549	0.878	-	-	-
gpt2	remiplus	note-on/off	CE	70	5	1.2	1.904	2.542	0.860	-	-	-
gpt2	remiplus	note-on/off	CE	90	5	1.2	1.817	2.529	0.879	-	-	-
gpt2	remiplus	note-on/off	CE	100	5	1.2	1.861	2.513	0.832	-	-	-
gpt2	midilike	time-shift	CE	10	5	1.2	1.474	2.279	0.788	-	-	-
gpt2	midilike	time-shift	CE	30	5	1.2	1.824	2.612	0.861	-	-	-
gpt2	midilike	time-shift	CE	50	5	1.2	1.643	2.547	0.889	-	-	-
gpt2	midilike	time-shift	CE	70	5	1.2	1.885	2.632	0.871	-	-	-
gpt2	midilike	time-shift	CE	90	5	1.2	1.795	2.490	0.879	-	-	-
gpt2	midilike	time-shift	CE	100	5	1.2	1.824	2.483	0.857	-	-	-
gpt2	remi	bar-based	CE	10	5	1.2	1.805	2.512	0.743	-	-	-
gpt2	remi	bar-based	CE	30	5	1.2	1.944	2.624	0.823	-	-	-
gpt2	remi	bar-based	CE	50	5	1.2	1.747	2.505	0.870	-	-	-
gpt2	remi	bar-based	CE	70	5	1.2	1.817	2.491	0.865	-	-	-
gpt2	remi	bar-based	CE	90	5	1.2	1.792	2.479	0.879	-	-	-
gpt2	remi	bar-based	CE	100	5	1.2	1.768	2.452	0.883	-	-	-

All Experiment Results (Continued)

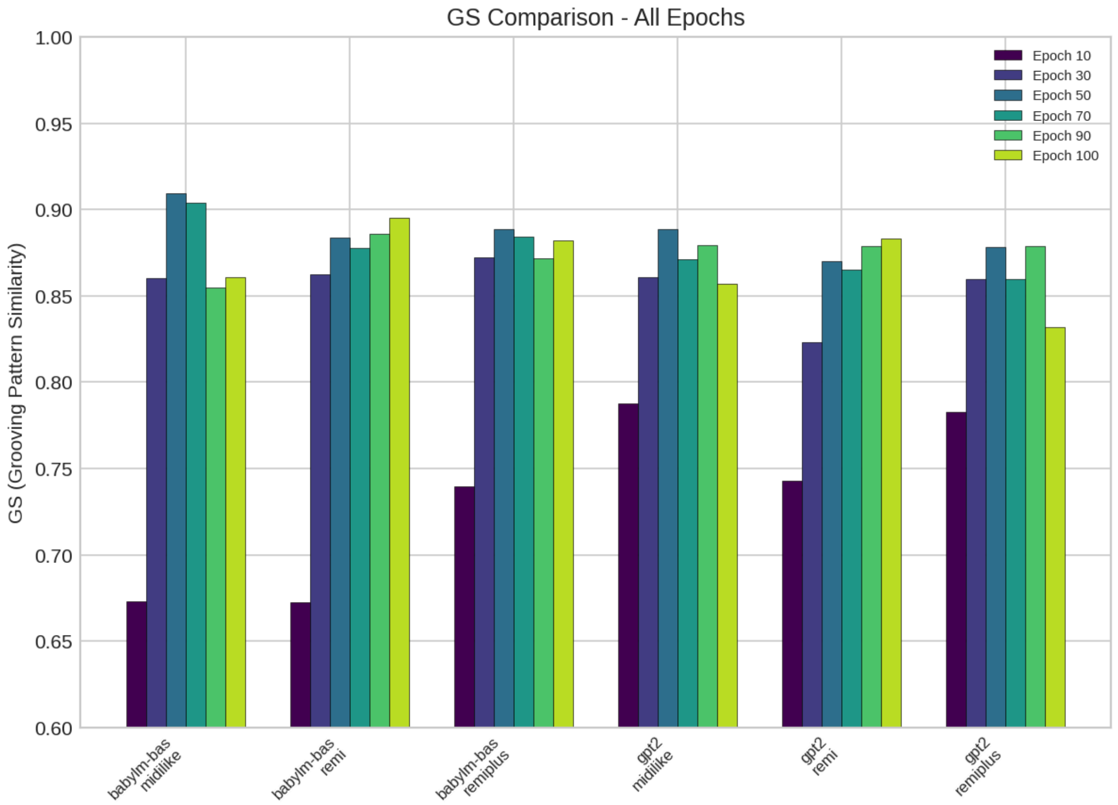
Model	Representation	Event	Loss	Epoch	Top-k	Temp	H1	H4	GS	SI_short	SI_mid	SI_long
babylm-gpt2	remi	bar-based	CE	10	5	1.2	2.048	2.652	0.673	-	-	-
babylm-gpt2	remi	bar-based	CE	30	5	1.2	1.902	2.665	0.862	-	-	-
babylm-gpt2	remi	bar-based	CE	50	5	1.2	1.859	2.637	0.884	-	-	-
babylm-gpt2	remi	bar-based	CE	70	5	1.2	1.728	2.459	0.878	-	-	-
babylm-gpt2	remi	bar-based	CE	90	5	1.2	1.763	2.428	0.886	-	-	-
babylm-gpt2	remi	bar-based	CE	100	5	1.2	1.785	2.483	0.895	-	-	-
babylm-gpt2	midilike	time-shift	CE	10	5	1.2	2.090	2.753	0.673	-	-	-
babylm-gpt2	midilike	time-shift	CE	30	5	1.2	1.875	2.704	0.860	-	-	-
babylm-gpt2	midilike	time-shift	CE	50	5	1.2	1.739	2.624	**0.909**	-	-	-
babylm-gpt2	midilike	time-shift	CE	70	5	1.2	1.851	2.514	0.904	-	-	-
babylm-gpt2	midilike	time-shift	CE	90	5	1.2	1.842	2.482	0.855	-	-	-
babylm-gpt2	midilike	time-shift	CE	100	5	1.2	1.805	2.494	0.861	-	-	-
babylm-gpt2	remiplus	note-on/off	CE	10	5	1.2	1.924	2.668	0.739	-	-	-
babylm-gpt2	remiplus	note-on/off	CE	30	5	1.2	1.807	2.657	0.872	-	-	-
babylm-gpt2	remiplus	note-on/off	CE	50	5	1.2	1.801	2.585	0.889	-	-	-
babylm-gpt2	remiplus	note-on/off	CE	70	5	1.2	1.874	2.493	0.884	-	-	-
babylm-gpt2	remiplus	note-on/off	CE	90	5	1.2	1.744	2.435	0.872	-	-	-
babylm-gpt2	remiplus	note-on/off	CE	100	5	1.2	1.830	2.476	0.882	-	-	-

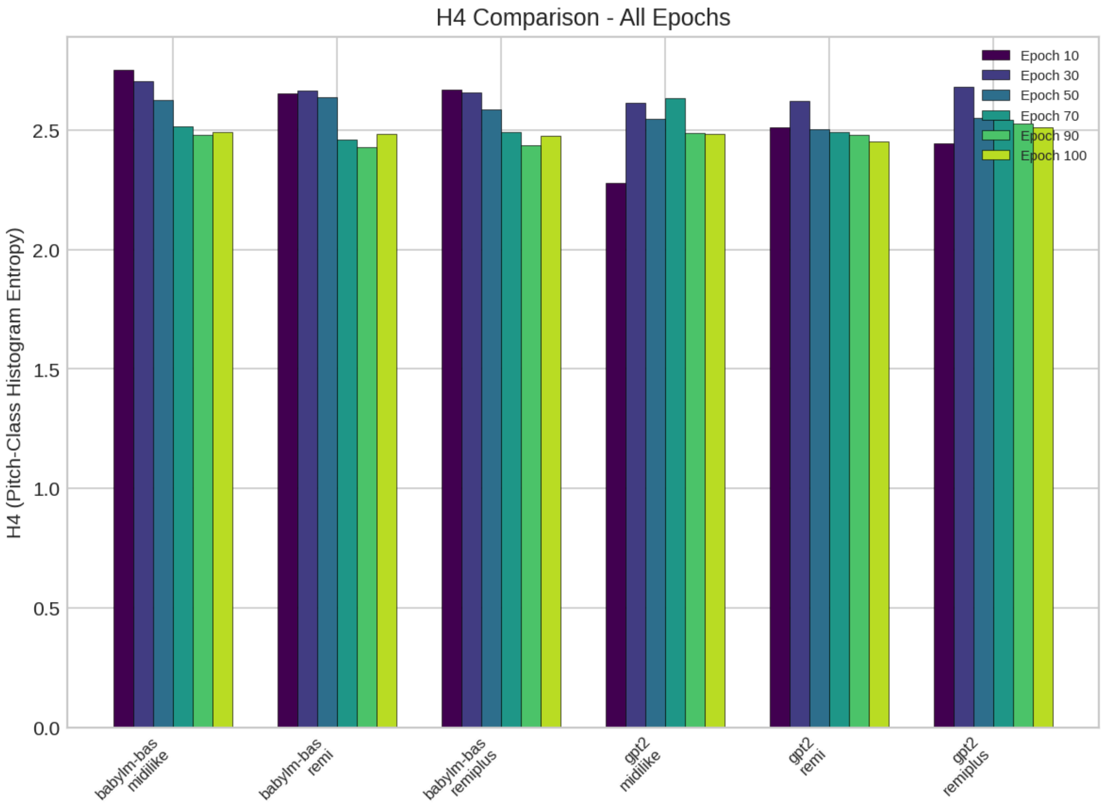
Model vs Tokenizer Heatmap



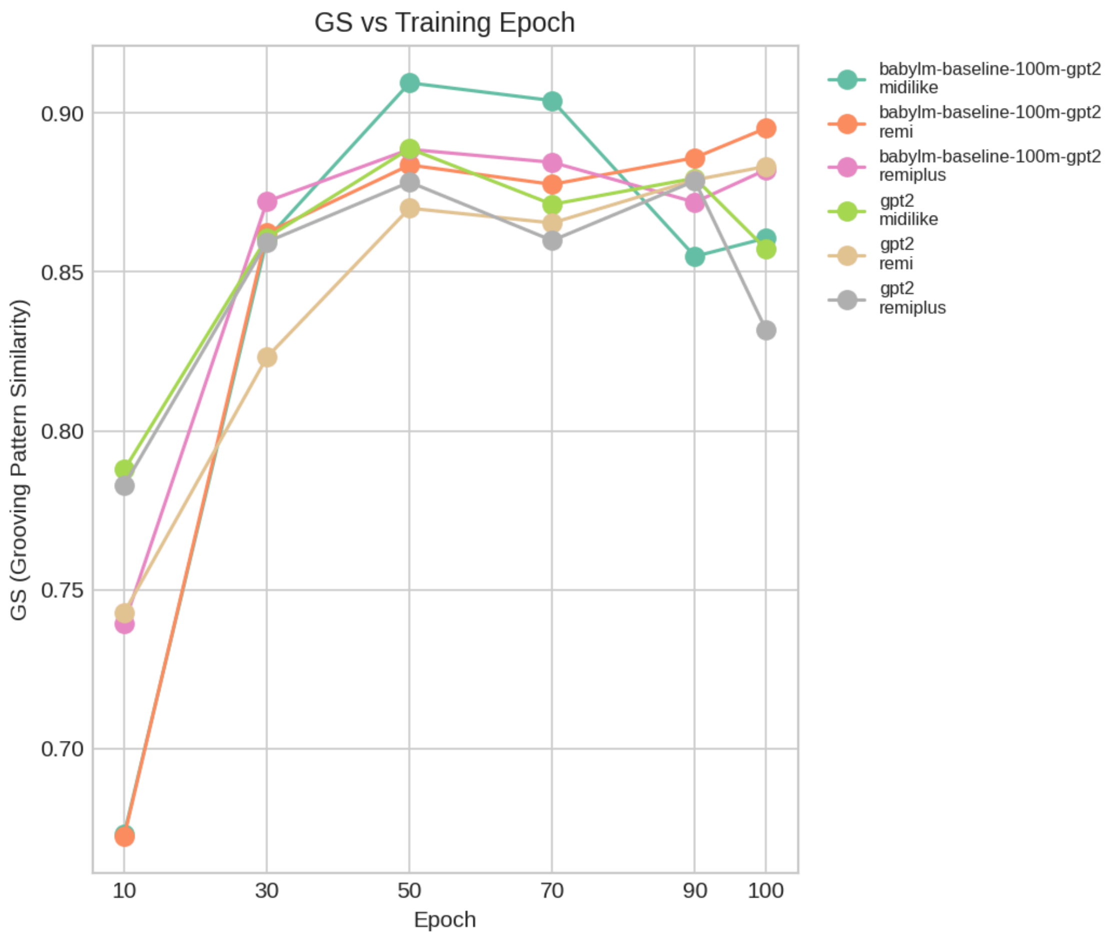


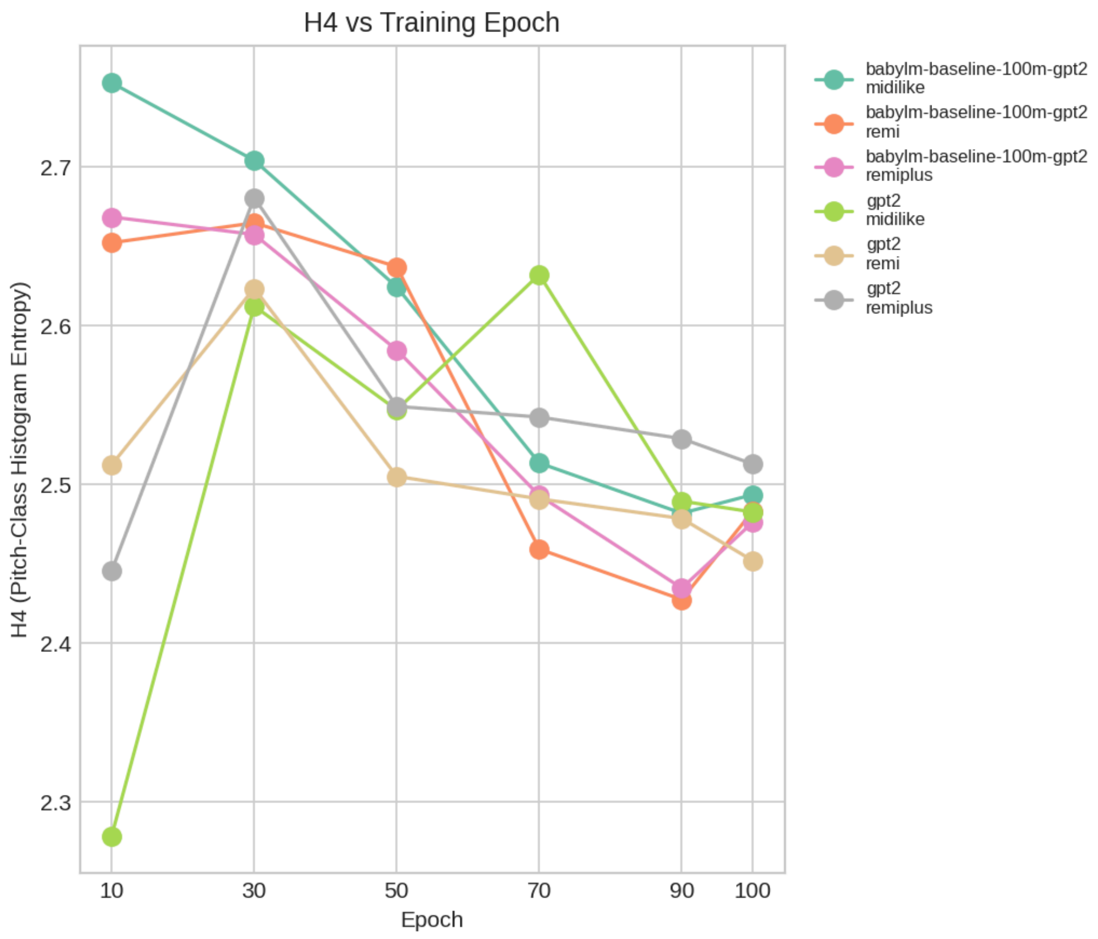
Model Comparison



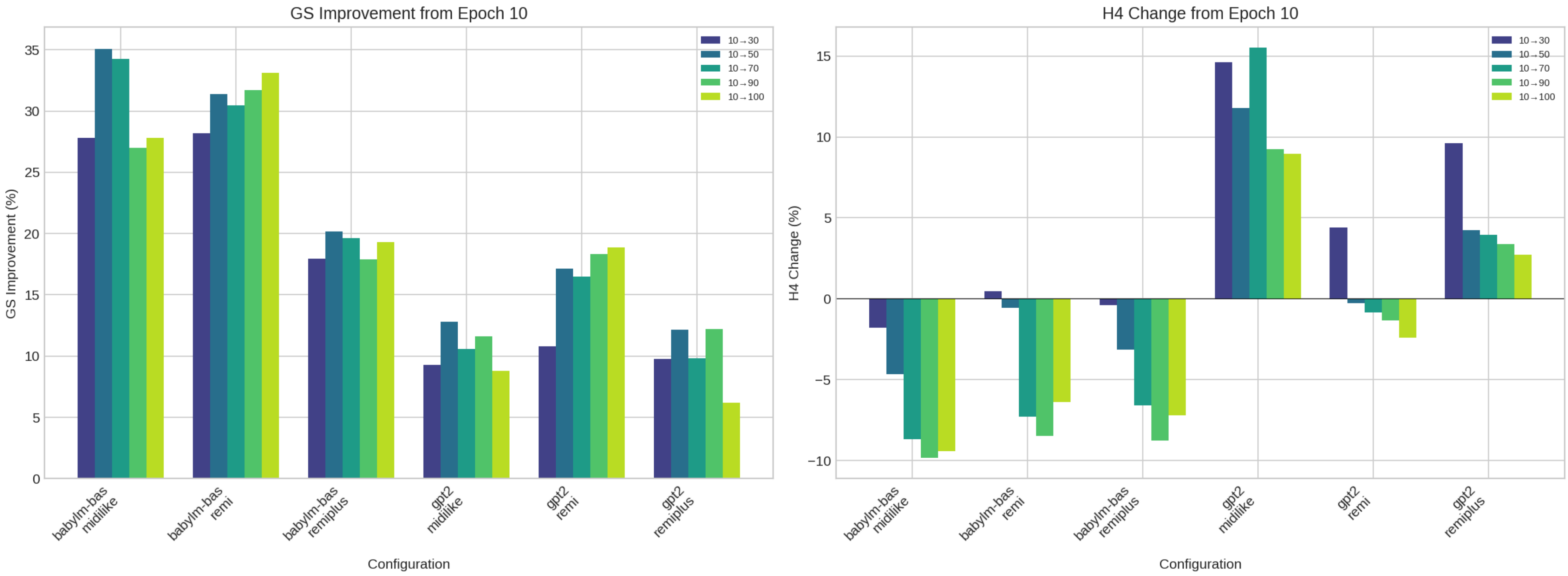


Training Progress - GS and H4 vs Epoch





GS and H4 Change from Epoch 10



Discussion: Controlled Variable Analysis

Effect of Training Epochs

Fixed: Model (BabyLM-GPT2), Tokenizer (REMI)

Findings:

- GS improves dramatically +28% from epoch 10→30
- Model learns rhythmic patterns quickly in early training
- Convergence plateau after epoch 50 (diminishing returns)
- H4 decreases at later epochs → model becomes more conservative in pitch choices
- Sweet spot: epoch 90-100 for best GS/H4 balance

Epoch	H4	GS
10	2.652	0.673
30	2.665	0.862
50	2.637	0.884
70	2.459	0.878
90	2.428	0.886
100	2.483	0.895

Effect of Model Architecture

Fixed: Tokenizer (REMI), Epoch (100)

Findings:

- BabyLM-GPT2 outperforms GPT-2 by +1.4% GS
- Pre-trained weights transfer well to music domain
- Language model pre-training provides better sequence understanding
- H4 nearly identical (~2.47) → architecture doesn't affect pitch diversity
- **Insight:** Pre-training on text helps model learn sequential patterns applicable to music

Model	H4	GS
GPT-2	2.452	0.883
BabyLM-GPT2	2.483	0.895

Effect of Tokenization (BabyLM-GPT2)

Fixed: Model (BabyLM-GPT2), Epoch (100)

Findings:

- REMI achieves best GS (0.895)
- REMI's explicit bar markers help maintain rhythmic structure
- MIDILike's event-based encoding loses temporal hierarchy
- H4 similar (~2.48) → tokenization doesn't affect pitch diversity
- **Insight:** Explicit structural tokens (bars) crucial for rhythm consistency

Tokenizer	H4	GS
REMI	2.483	0.895
REMI+	2.476	0.882
MIDILike	2.494	0.861

Effect of Tokenization (GPT-2)

Fixed: Model (GPT-2), Epoch (100)

Findings:

- Same pattern confirmed: REMI best for both models
- REMI+ shows worst GS (0.832) - chord tokens may add noise
- Effect consistent across architectures → tokenization more important than model choice
- **Insight:** REMI+ chord tokens may confuse the model, hurting rhythm more than helping harmony

Tokenizer	H4	GS
REMI	2.452	0.883
REMI+	2.513	0.832
MIDILike	2.483	0.857

Conclusion - Task 1

Top 3 by GS

Rank	Model	Tokenizer	Epoch	GS
1	BabyLM-GPT2	MIDILike	50	**0.909**
2	BabyLM-GPT2	MIDILike	70	0.904
3	BabyLM-GPT2	REMI	100	0.895

Top 3 by H4

Rank	Model	Tokenizer	Epoch	H4
1	BabyLM-GPT2	MIDILike	10	**2.753**
2	BabyLM-GPT2	MIDILike	30	2.704
3	GPT-2	REMI+	30	2.681

Final Model Selection

I choose BabyLM-GPT2 | REMI | 100 for submission.

Metric	Value
H1	1.785
H4	2.483
GS	**0.895**

Reasoning:

- **Best listening quality**: Subjectively sounds the most musical and coherent among all configurations
- **Best balance**: GS of 0.895 indicates strong rhythmic consistency while maintaining reasonable melodic diversity (H4 = 2.483)

Key Insights

1. **GS vs H4 trade-off**: High H4 occurs at early epochs with low GS
   - Best GS at epoch 50-70, best H4 at epoch 10-30
2. **BabyLM-GPT2 dominates**: All top 3 GS and 2/3 top H4 use BabyLM-GPT2
   - Pre-training transfers effectively to music generation
3. **MIDILike peaks early**: Best performance at intermediate epochs (50-70)
   - May overfit at later epochs
4. **REMI stable at epoch 100**: Consistent performance without degradation

Task 2: Conditional Generation

Overview

Generate 24-bar continuations from 8-bar MIDI prompts provided by TA.

Output: 32 bars total (8 prompt + 24 generated)

Experiment Design

Checkpoints Tested

Based on Task 1 results, I manually selected 3 promising checkpoints:

Checkpoint	Tokenizer	Epoch
1	REMI	90
2	MIDILike	100
3	REMI	100

All using **BabyLM-GPT2** model.

Inference Settings

Each checkpoint tested with 4 configurations:

Setting	Top-k	Temp	Rep. Penalty
Default	5	1.2	1.2
Balanced	50	1.0	1.2
Creative	100	1.5	1.1
Conservative	20	0.8	1.3

Total: 3 checkpoints × 4 settings × 3 songs = 36 generations

Best Continuation Selection

For each prompt song, the best result selected by manual listening:

Song	Best Checkpoint	Tokenizer	Epoch	Setting
Song 1	Checkpoint 1	REMI	90	Default
Song 2	Checkpoint 2	MIDILike	100	Default
Song 3	Checkpoint 3	REMI	100	Default

Note: All best results used the **Default** setting (top-k=5, temp=1.2). Lower top-k produces more coherent continuations.

Observations

Song	Checkpoint	Performance	Notes
Song 1	REMI (90)	Best	Successfully captured song patterns and rhythm
Song 2	MIDILike (100)	Worst	Complex prompt song made continuation difficult
Song 3	REMI (100)	Moderate	Partially captured scale patterns

Conclusion - Task 2

1. **Different checkpoints work best for different prompts**
   - No single configuration is universally optimal
   - Tokenizer choice significantly affects continuation style
2. **REMI performs better for higher BPM songs**
   - Bar-based structure handles faster tempos more effectively
   - MIDILike better suited for moderate tempo pieces
3. **Top-k may have significant impact on generation quality**
   - All best results used the Default setting with top-k=5
   - Lower top-k possibly produces more coherent continuations
   - This hypothesis requires further investigation in future work

References

- **GPT-2**: Radford, A., et al. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI*.
- **BabyLM**: Warstadt, A., et al. (2023). Findings of the BabyLM Challenge. *CoNLL*.
- **MidiTok**: Fradet, N., et al. (2021). MidiTok: A Python package for MIDI file tokenization. *ISMIR*.
- **REMI**: Huang, Y.-S., & Yang, Y.-H. (2020). Pop Music Transformer. *ACM MM*.
- **MusDr**: Wu, S., & Yang, Y.-H. (2020). The Jazz Transformer on the Front Line. *ISMIR*.
- **PyTorch**: Paszke, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *NeurIPS*.
- **Reference Implementation**: Liao, J.-W. (2024). Symbolic Music Generation.

Thank you for your time