Mo Bamba:
A Mamba LLM For Rap Generation

Dylan Lu

1 Introduction

Last week, IBM released Bamba[1]—a Mamba-based LLM that outperforms SoTA while requiring 5× less training data and delivering 2.5× faster inference. Inspired by this recent success, I tested Mamba's capabilities by finetuning it on a rap lyric dataset, evaluating its performance against traditional transformer LLMs.

2 Methods

Using a 40GB A100 GPU, my approach is as follows: 1) Download 3B versions of Mamba 2 and Llama 3.2 from Huggingface, 2) Download a dataset containing 1k+ lyrics from Genius.com, 3) Train using both LoRA, full finetuning, and 4-bit QLoRA (Llama only). I trained for 3 epochs with a sequence length of 2048, batch size 2, and learning rate 2e-4.

3 Evaluation

To score my models, I use Microsoft's G-Eval[2]—a GPT-based evaluator that averages scores on coherence, consistency, fluency, and relevance (1–5). I start generation with the prompt: ”[Verse 1]\n”.

Mamba 2 (3B)	**G-Eval**	**Time (min)**
Base Model	2.00	-
Full Finetune	3.25	49.15
LoRA (r=16)	4.00	43.25
LoRA (r=8)	3.50	41.68

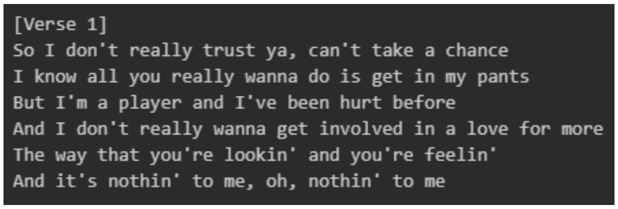
Llama 3.2 (3B)	**G-Eval**	**Time (min)**
Base Model	2.00	-
Full Finetune	3.50	32.21
LoRA (r=16)	3.50	16.32
LoRA (r=8)	3.50	14.57
QLoRA (r=16)	3.75	21.07

4 Results

Mamba performs quite similarly to Llama, despite being pretrained on 3× less training data. Empirically, both models start off very well, but they eventually end up infinitely repeating some phrases. This issue likely stems from the fact that many songs use repetition in certain verses, causing the model to assign high probability to continuing the pattern once it appears.

As for differences between training methods, LoRA and QLoRA perform on par with full finetuning, though are faster and less memory-intensive. Decreasing the LoRA rank does not seem to hurt performance as well. For Llama, I use Unsloth[3]—which wraps Huggingface with fused Triton kernels and other optimizations—to gain additional speedup with no loss of accuracy. Mamba is not yet supported by Unsloth, which is why the speed benefit more apparent for Llama.

4.1 Sample Output (Mamba 3B)



```
[Verse 1]
So I don't really trust ya, can't take a chance
I know all you really wanna do is get in my pants
But I'm a player and I've been hurt before
And I don't really wanna get involved in a love for more
The way that you're lookin' and you're feelin'
And it's nothin' to me, oh, nothin' to me
```

5 Lessons Learned

I learned how to set up finetuning pipelines using Huggingface and Unsloth, making it easy to experiment with different base models. I realized that evaluating creative tasks like rap generation is difficult, since large-scale human feedback isn't practical. For future projects, I look forward to applying LoRA as a lightweight and effective alternative to full finetuning.

[1] Tri Dao and IBM 2025.

[2] Liu et al. 2023.

[3] Daniel Han and team 2023.