

Fine-Tuning de Modèle de Langage pour la Génération Musicale

Groupe MAPo7
17 mai 2024



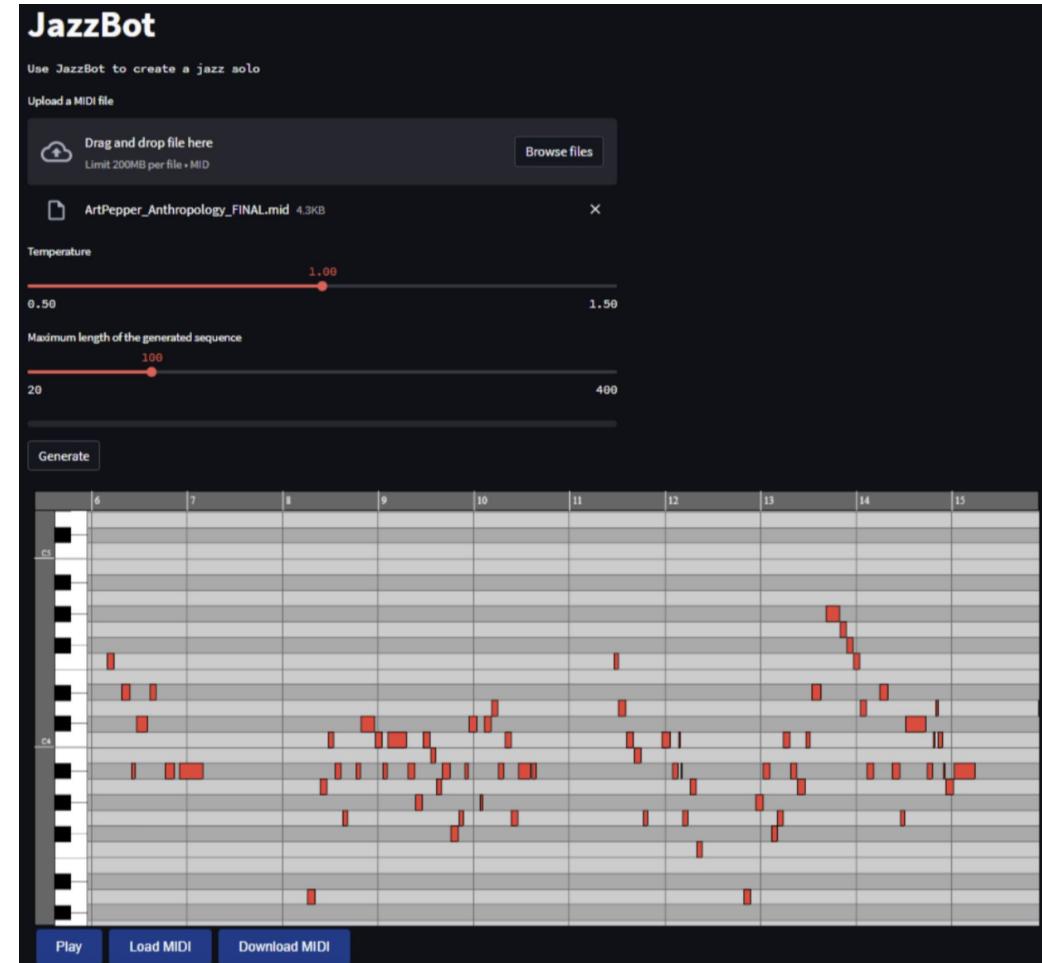
Introduction

- Essor phénoménal de l'IA générative
- Les *Large Language Model* (LLM) ne font que prédire le prochain *token*:
 - perroquet stochastique ou *compréhension* de l'harmonie et du rythme ?
 - une tâche créative
- Notre ambition : une assistance au compositeur



Introduction

- Reprise du sujet d'un groupe d'X 21
- Changement radical d'approche :
 - *training from scratch vs fine-tuning*
 - même ordre de grandeur du nombre de paramètres entraînés.
- Choix de ne pas faire d'application intégrée pour se concentrer sur la génération.



Sommaire

I - Constitution du Dataset

- a) Première approche
- b) Transcription en MIDI et Data Augmentation
- c) La notation Simplified MIDI

II- Entraînement du modèle

- a) Architecture des Large Language Model (LLM)
- b) Les couches d'optimisation et nos améliorations successives

III-Résultats et limitations

I- Constitution du Dataset

a) Première approche

- émuler le fonctionnement d'une base de donnée reconnue en *Music Information Retrieval* (MIR) : le *web-scraping* de *Giant MIDI Piano*.
- Un côté brouillon inhérent au *scraping*
- ré-orientation vers la transcription de bases de donnée existante : *DadaGP* dans le format *Guitar Pro*.



Figure 3: Audio de guitare traité par un convertisseur naïf (l'audio original est un thème à une seule voix) : Les mêmes effets que pour le piano sont présents, avec une démultiplication des notes, tant par l'ajout des harmoniques que par leur fragmentation temporelle.

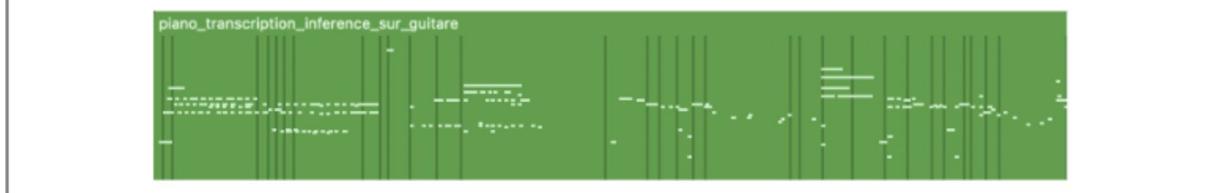


Figure 4: Audio de guitare traité par un convertisseur entraîné sur le piano : L'entraînement diminue beaucoup la tendance à démultiplier les harmoniques, cependant, on ne la supprime pas totalement (rappelons que l'audio consiste en une seule voie). Au niveau de la fragmentation temporelle cependant, on ne fait que très peu de progrès. Le convertisseur marche moins bien que sur du piano.

I- Constitution du Dataset

Format Guitar Pro

The image shows a screenshot of a Guitar Pro file for the song "Little Wing" by Jimi Hendrix. The title "LITTLE WING" is at the top, followed by "Jimi Hendrix" and "Axis: Bold As Love". The key signature is G major (one sharp). The tempo is "Slow Ballad" with a tempo of 68 BPM. The guitar tablature shows six tuning pegs at the top: E, B, G, D, A, E. The music starts with a piano intro, followed by a guitar solo section labeled "let ring" three times, ending with a "F.M." (Feedback Modulation) effect. The guitar part includes various techniques like hammer-ons, pull-offs, and bends.

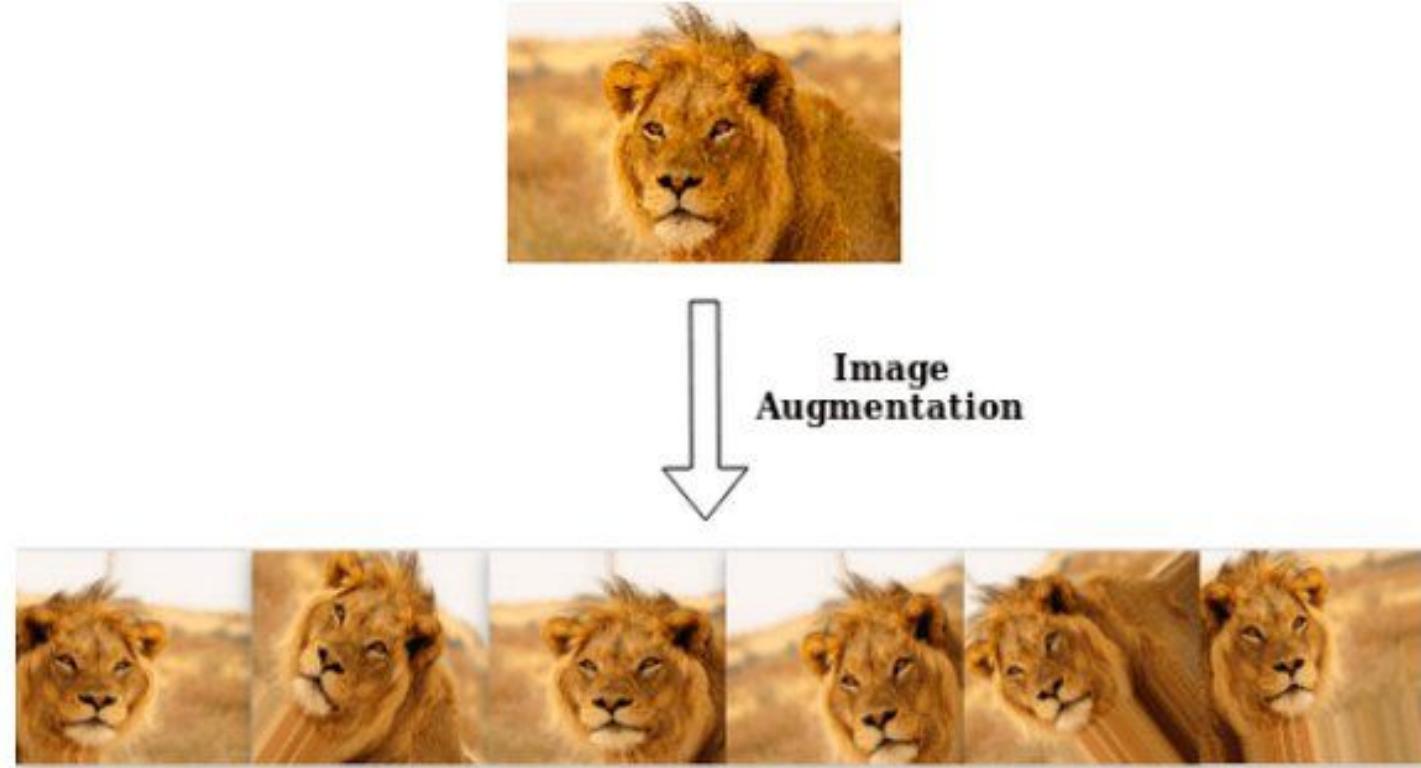
Sélection des artistes de metal

- > Mudvayne
- > Mourning Beloveth
- > Motorhead
- > Motley Crue
- > Morbid Angel
- > Moore, Vinnie
- > Moonspell
- > Misanthrope
- > Mindslave
- > Michael Schenker Group
- > Michael Angelo
- > Metallica
- > Metalium
- > Metal Church
- > Meshuggah
- > Mercyful Fate
- > Merauder
- > Megadeth
- > Mayhem
- > Mastodon
- > Masterplan

Source : <https://cameronrwolfe.substack.com/>

I- Constitution du Dataset

b) Data Augmentation



Partition du morceau original (Creeping Death, Metallica)

Synthétiseur de dents de scie, Voice

Guitare électrique, Guitar

Creeping Death

$\text{♩} = 183$

The musical score for the original 'Creeping Death' by Metallica. The top staff, labeled 'Synthétiseur de dents de scie, Voice', features a single note on each of the five lines of the treble clef staff. The bottom staff, labeled 'Guitare électrique, Guitar', shows a sixteenth-note pattern followed by a sustained note. The tempo is marked as quarter note = 183.

Synthétiseur de dents de scie, Voice

Guitare électrique, Guitar

Creeping Death

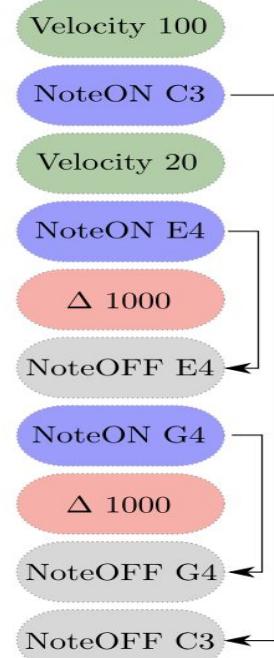
$\text{♩} = 133$

The musical score for the modified version of 'Creeping Death'. The top staff, labeled 'Synthétiseur de dents de scie, Voice', features a single note on each of the five lines of the treble clef staff. The bottom staff, labeled 'Guitare électrique, Guitar', shows a melodic line with eighth-note pairs and sustained notes. The tempo is marked as quarter note = 133.

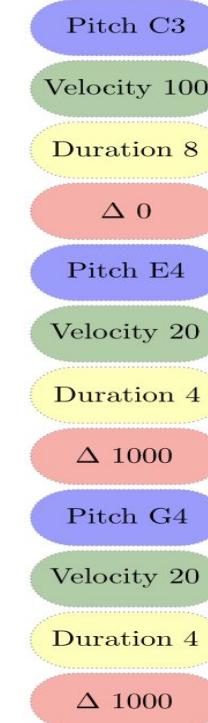
Partition du morceau modifié

I- Constitution du Dataset

c) La notation Simplified MIDI



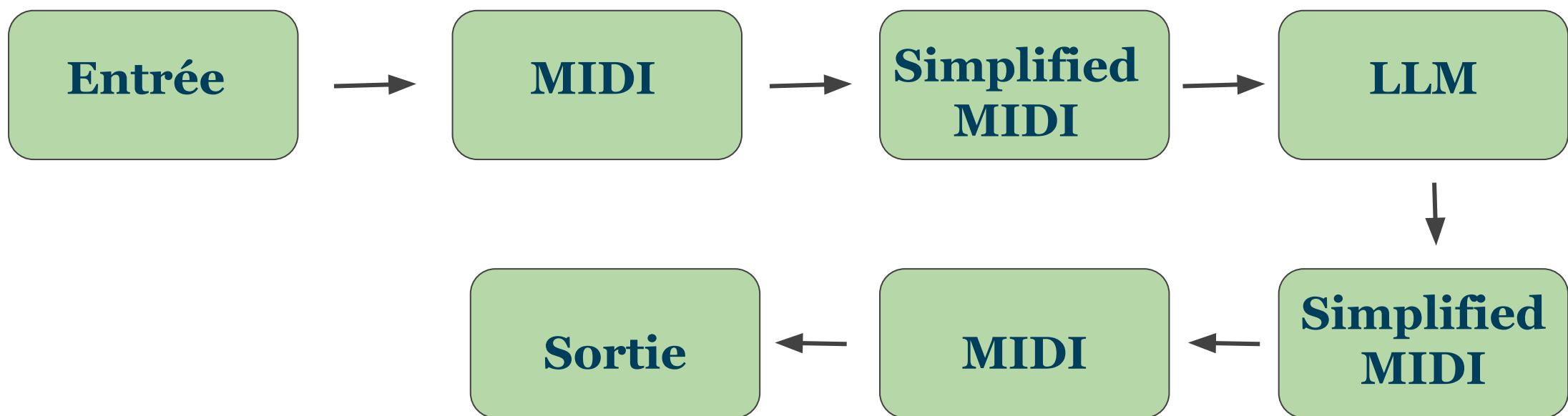
MIDI



Simplified MIDI

I- Constitution du Dataset

c) La notation Simplified MIDI



=> Simplified MIDI

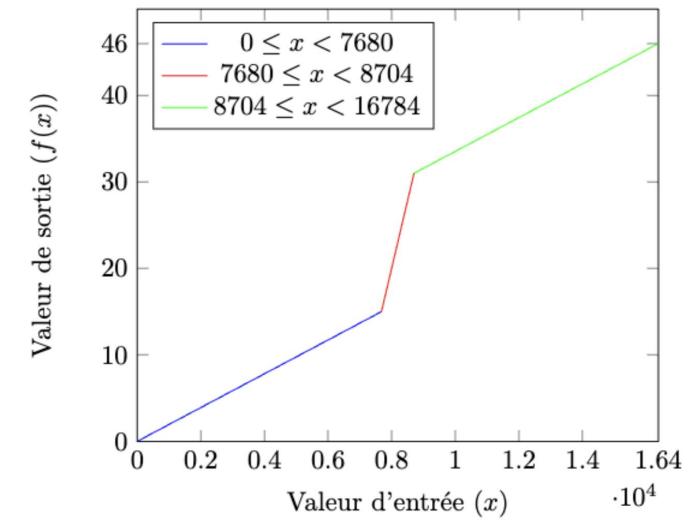
Pitch, velocity, duration,
time

I- Constitution du Dataset



Le Pitch Bend

- Ajouté au Simplified MIDI
- Nécessaire pour reproduire fidèlement la guitare
- Diminution de la précision pour l'apprentissage



Exemple en Zero Shot avec ChatGPT4

You are a classical pianist composer. In this context, each music note in a musical sequence is described using four parameters: pitch (p) from 0 to 127 (highest pitch), volume (v) from 0 to 127 (loudest), duration of the note (d) in ticks, and the length of the pause (t) in ticks before the next note begins regardless of the previous note's duration. A tick is approximately 5.21 milliseconds. Each parameter is followed by its value and separated by colons (e.g. p52:v57:d195:t212). Your composition should demonstrate a clear progression and development, appropriate pauses, including thoughtful variations in melody, harmony, rhythm.

Your Task is to complete the generation of : p45:v73:d389:t1 p84:v84:d388:t7
p76:v62:d227:t0 p48:v63:d5:t110 p52:v69:d394:t5 p81:v78:d389:t117



To continue and complete the musical sequence you provided with a thoughtful progression and variations, let's analyze the existing notes first:

1. **p45:v73:d389:t1** - A lower pitch (p45) played with medium volume (v73) and long duration (d389 ticks), followed by a very short pause (t1).
2. **p84:v84:d388:t7** - A high pitch (p84) played at a strong volume (v84) with almost the same long duration (d388 ticks) as the first, followed by a short pause (t7).
3. **p76:v62:d227:t0** - A mid-high pitch (p76) with quieter volume (v62) and a shorter duration (d227 ticks), with no pause before the next note.
4. **p48:v63:d5:t110** - A lower pitch (p48) close to the first note's pitch, played at a similar volume (v63) but extremely short duration (d5 ticks), followed by a longer pause (t110).
5. **p52:v69:d394:t5** - Another lower pitch (p52), played louder (v69) and with a long duration similar to the first and second notes, followed by a short pause (t5).
6. **p81:v78:d389:t117** - A high pitch (p81) similar to the second note, played loudly (v78) with a long duration (d389 ticks), ending with a longer pause (t117).

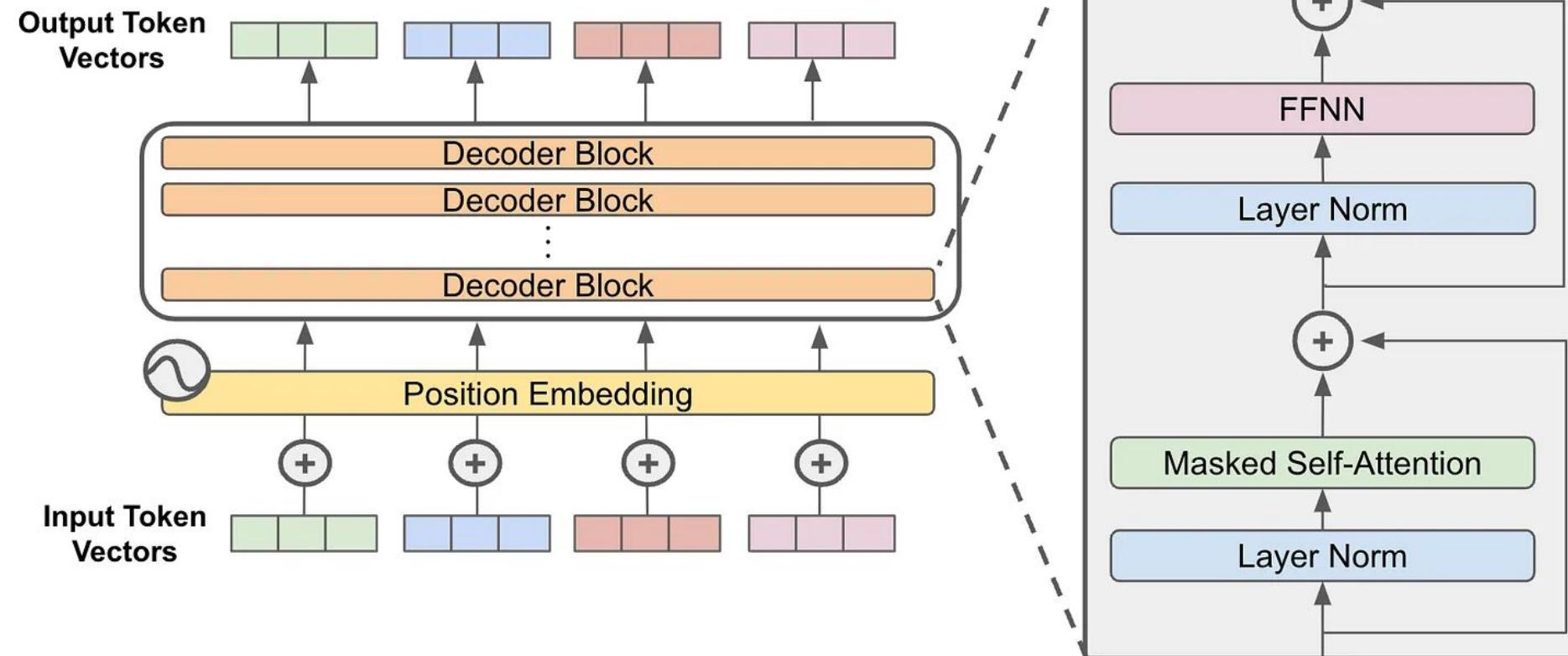
For a continuation that builds upon this and demonstrates a progression with variety in melody, harmony, and rhythm, here's a completion:

- **p53:v70:d250:t10** - Gradual ascent in pitch (p53), maintaining a medium-loud volume (v70) with a moderate duration (d250 ticks), followed by a pause (t10) to prepare for the next phrase.
- **p59:v75:d300:t15** - A slightly higher pitch (p59) with increased volume (v75), and a longer duration (d300 ticks) than the previous, with a moderate pause (t15) to enhance the melody's progression.
- **p60:v80:d120:t0** - Moving to a slightly higher pitch (p60) with a louder volume (v80) for a shorter, impactful note (d120 ticks), no pause to maintain flow.
- **p47:v65:d180:t100** - Return to a lower pitch (p47) with a slight decrease in volume (v65) for a medium duration (d180 ticks), followed by a significant pause (t100) to create a moment of reflection.
- **p68:v77:d350:t3** - A rise to a mid-high pitch (p68), louder (v77), with a long duration (d350 ticks), followed by a minimal pause (t3), pushing the momentum forward.
- **p55:v72:d400:t120** - Ending on a thoughtful note, a mid-range pitch (p55) with balanced volume (v72) and the longest duration in this extension (d400 ticks), followed by a significant pause (t120) to signify a conclusion.

This completion sequence provides a natural flow from the previous sequence, featuring dynamic shifts in volume and pitch while maintaining varied durations and thoughtful pauses that contribute to a cohesive and expressive musical piece.

II- Entraînement du Modèle

a) Architecture d'un LLM



Source : <https://cameronrwolfe.substack.com/>

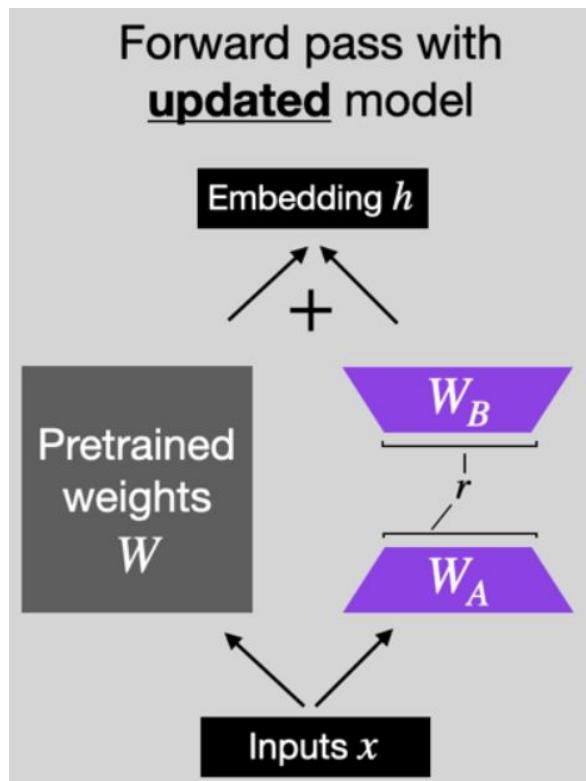
II- Entraînement du Modèle

La particularité du fine tuning des LLMs:

- taille/efficacité en temps et en mémoire
 - parallélisation
 - PEFT
- Optimisation des hyperparamètres

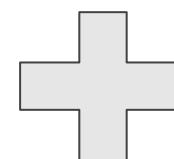
II- Entraînement du Modèle

b) Q-LoRA



Techniques de quantification

Quantification par blocs et par vecteurs



Un nouveau type de données :
NF4

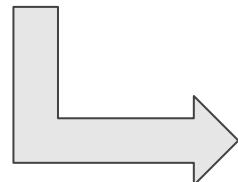
Source : <https://lightningaidev.wpeengine.com/>

Memory_usage = 34*seqlen*batch_size*hidden_size + 5*attention_heads*seqlen*seqlen*batch_size

II - Entraînement du modèle

c) Flash Attention 2

- Calculs de la matrice d'attention par blocs



Diminution des accès en écriture et lecture à la VRAM

- Parallélisation de ces calculs par blocs

Gain de 38,9% d'utilisation de VRAM et de 8,1% en temps de calcul



II- Entraînement du Modèle

c) Augmentation du contexte

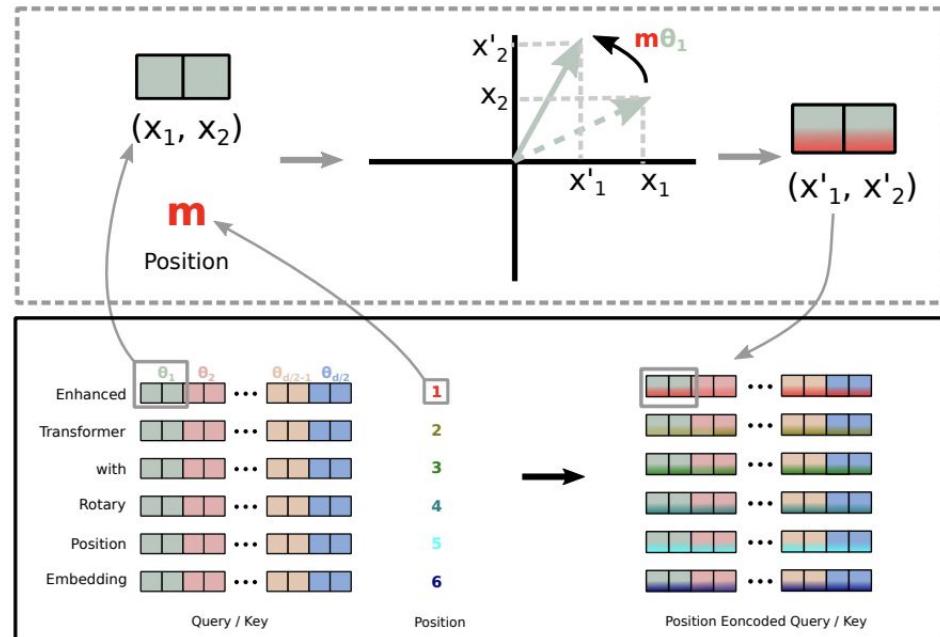


Figure 1: Implementation of Rotary Position Embedding(RoPE).

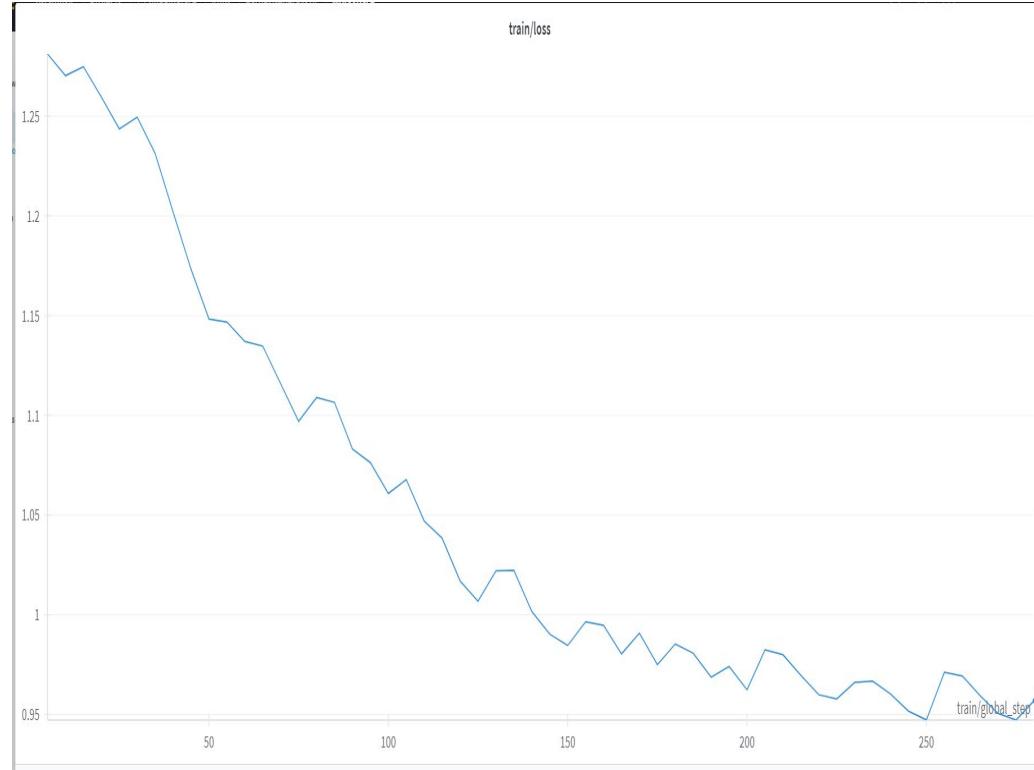
Contexte (en tokens)	Longueur des séquences d'entraînement
4096 (10 secondes)	4096
8192	8192
16384 (avec Adam 8bits)	16384
32768	8192

II- Entraînement du Modèle : Optimisation des hyperparamètres



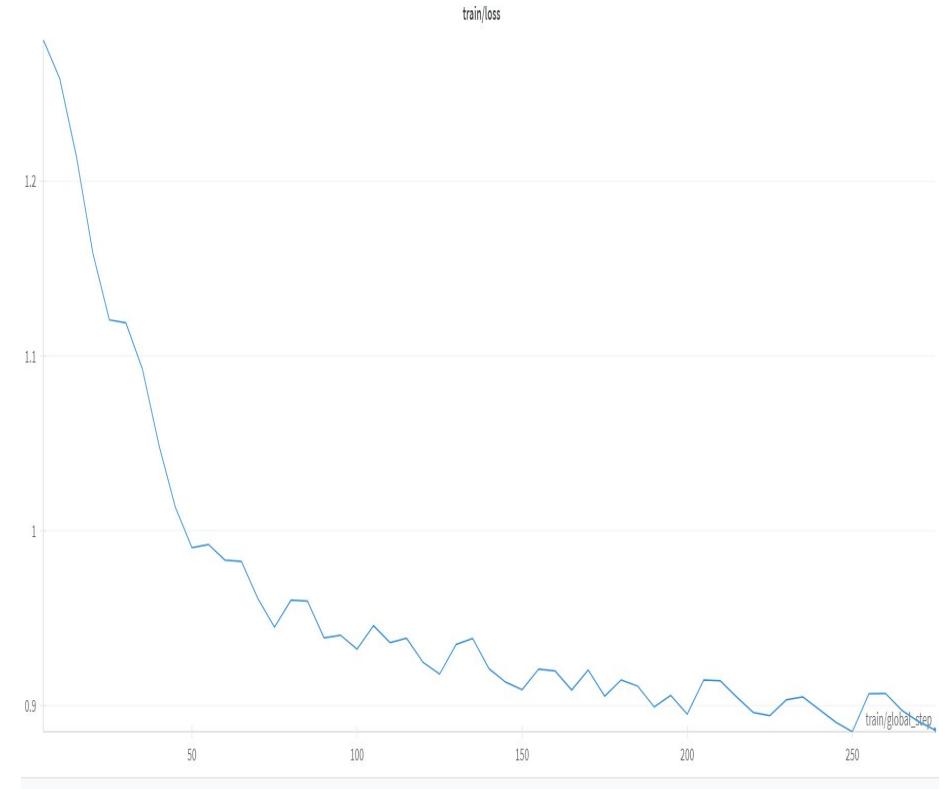
LEARNING RATE $2 \cdot 10^{-5} - 2 \cdot 10^{-4}$	BATCH SIZE GRADIENT ACCUMULATION	EPOCHS 3-10
LEARNING RATE SCHEDULER cosine, linear, constant, cosine with restarts	WARMUP 3%-10%	OPTIMIZER AdamW32bits, adafactor, Adam 8bits

II- Entraînement du Modèle : Learning rate



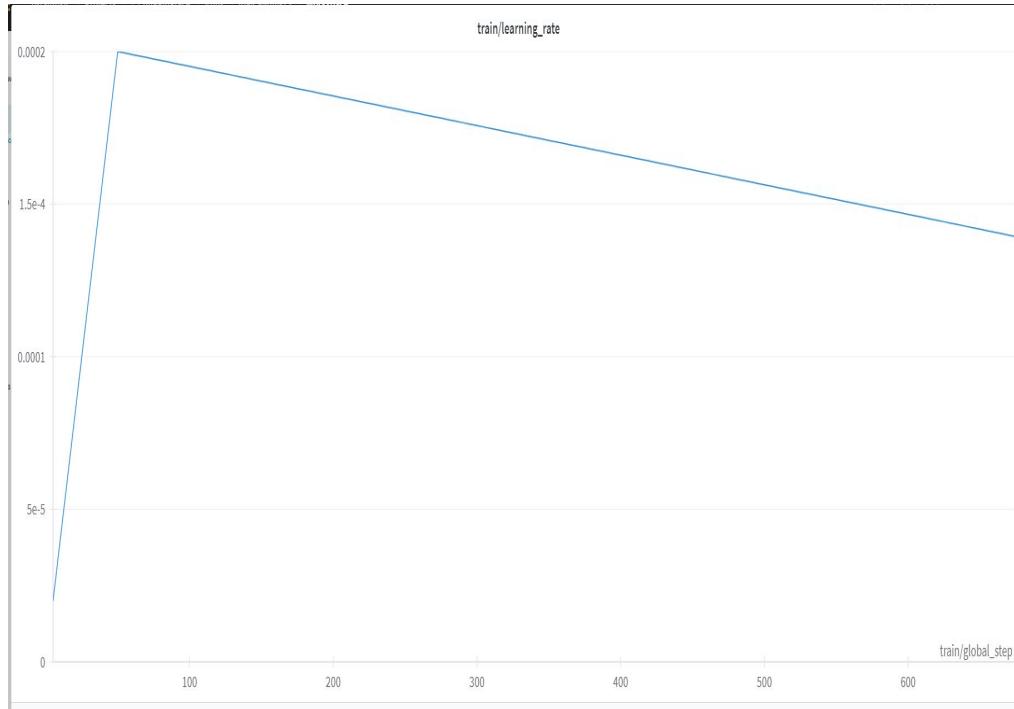
2 10 -5

Loss : 0.9569 vs 0.886 en 3h40

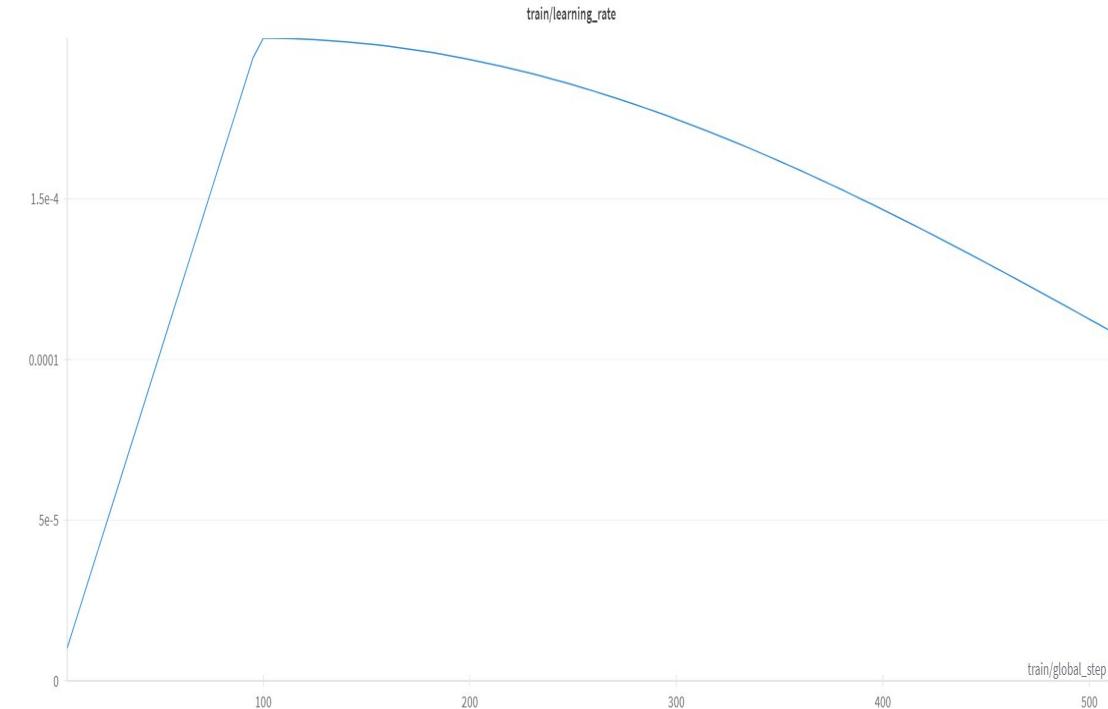


2 10 -4

II- Entraînement du Modèle : Scheduler



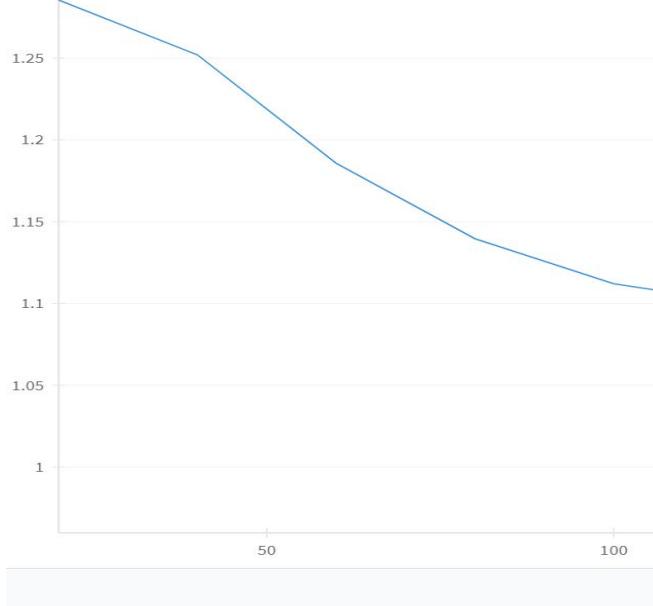
Linear



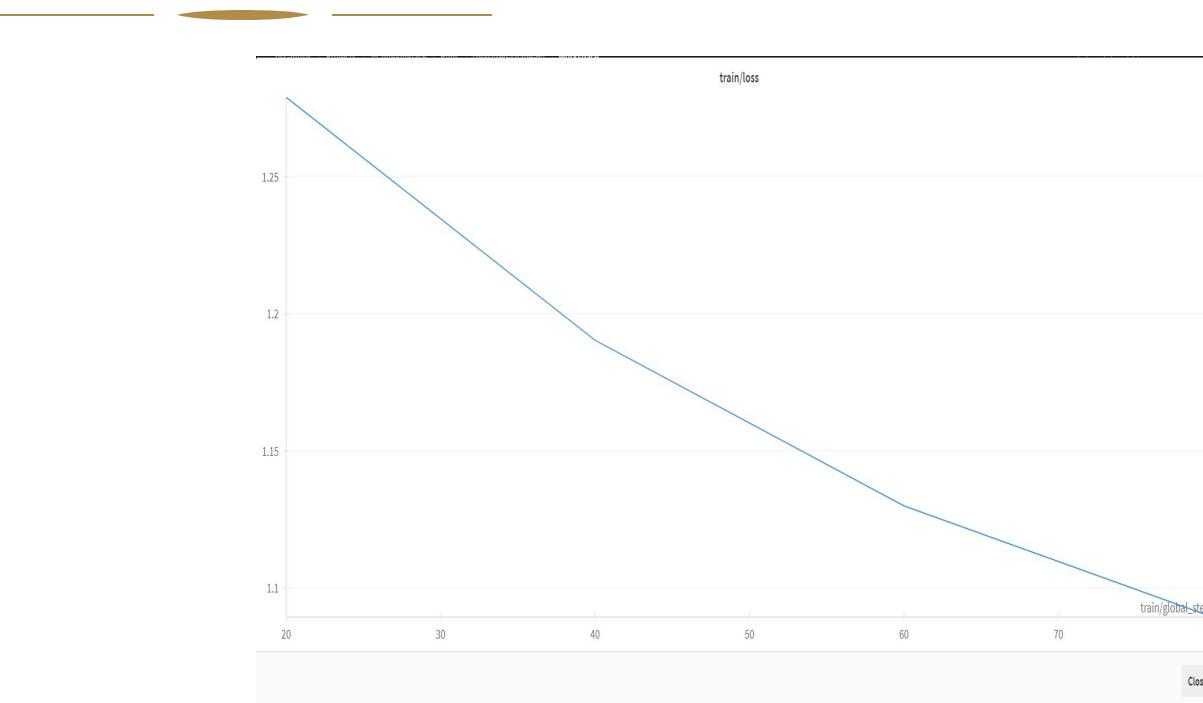
Cosine

Loss : 0.8716 vs 0.8609 en 7h30

II- Entraînement du Modèle : Warm up



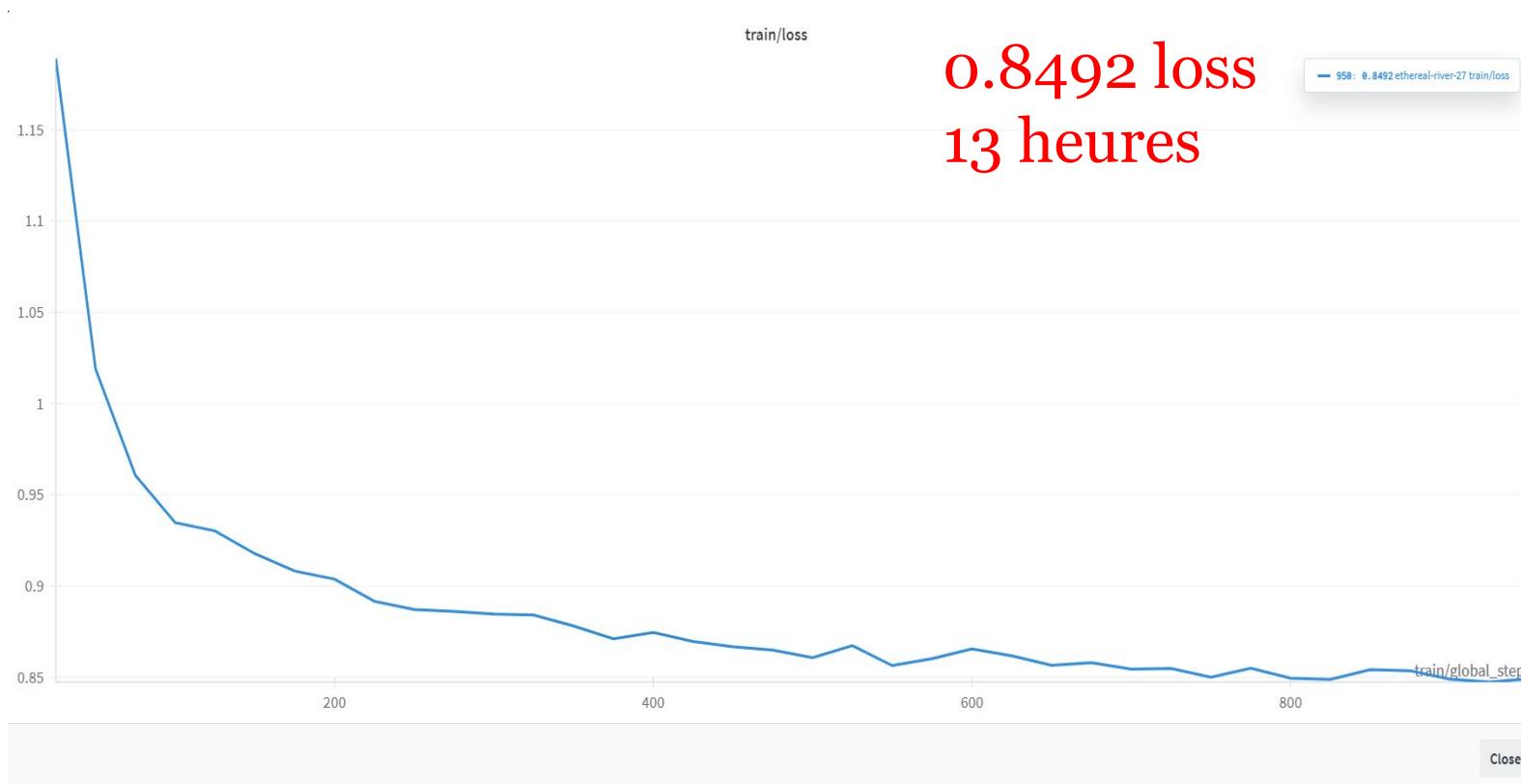
10%



3%

Loss : 1.1396 vs 1.0893 en 1h

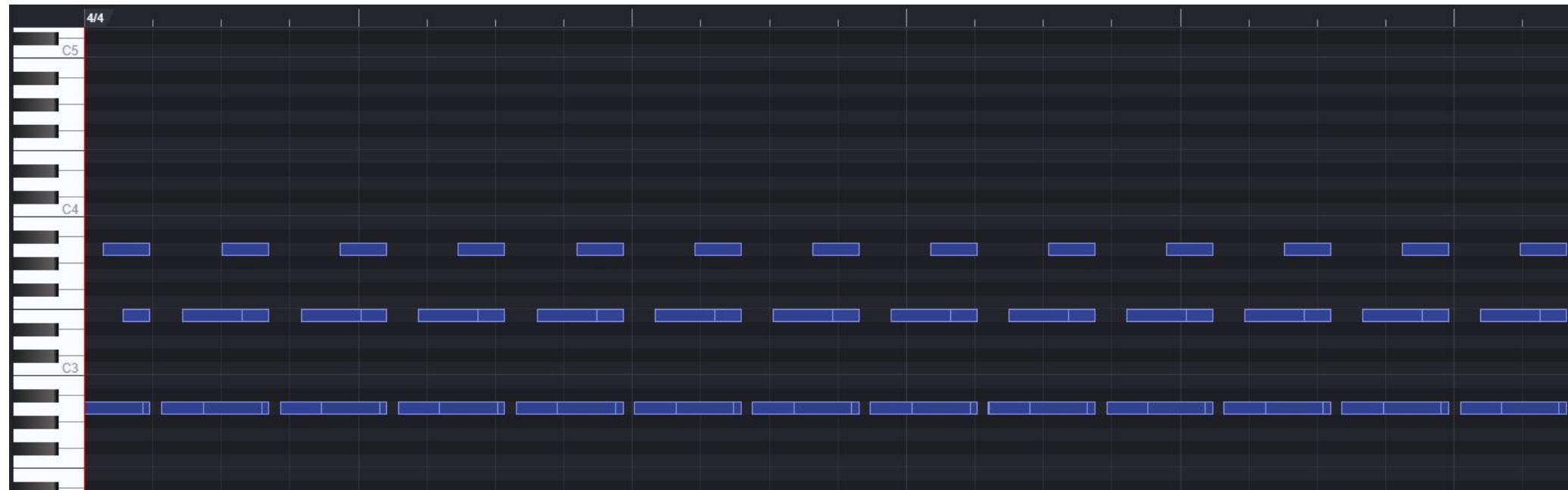
II- Entraînement du Modèle : Meilleur modèle



Epochs : 3
Learning rate : 2
10-4
Accumulation : 1
warmup ratio :
0.03
Lr_scheduler :
Cosine
Optim : AdamW

II- Entraînement du Modèle : Problème de répétition

Greedy search donne :



II- Entraînement du Modèle : Problème de répétition

Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation

Jin Xu^{1,*},Xiaojiang Liu⁴, Jianhao Yan², Deng Cai³, Huayang Li⁴, Jian Li¹

¹Institute for Interdisciplinary Information Sciences, Tsinghua University

²School of Engineering, Westlake University

³The Chinese University of Hong Kong

⁴Apple

A Theoretical Analysis of the Repetition Problem in Text Generation*

Zihao Fu,¹ Wai Lam,¹ Anthony Man-Cho So,¹ Bei Shi²

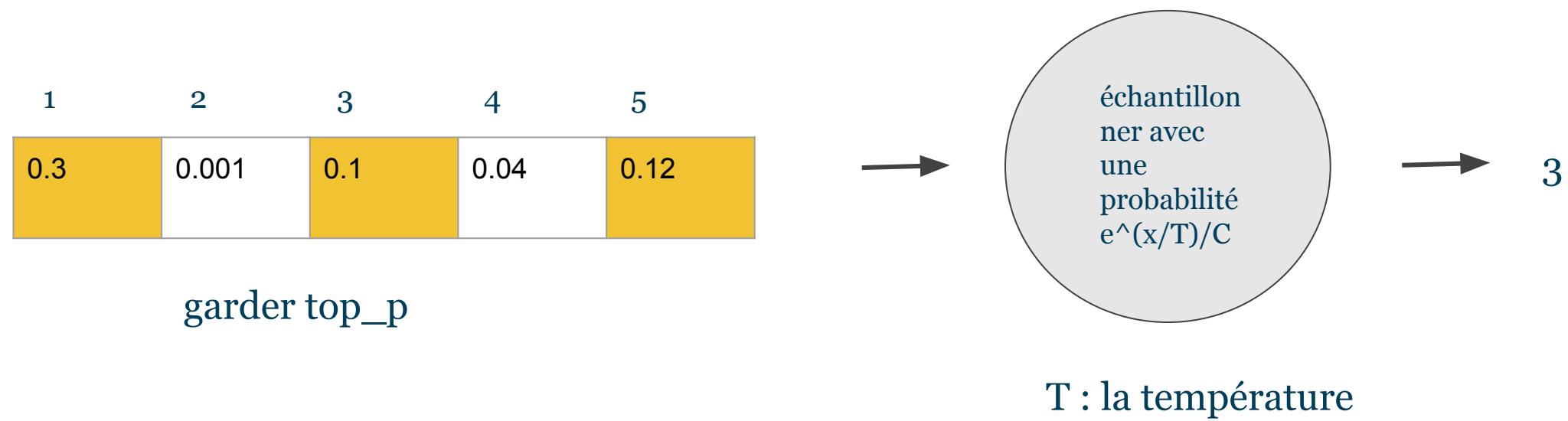
¹Department of Systems Engineering and Engineering Management,

The Chinese University of Hong Kong, Hong Kong

²AI Lab, Tencent

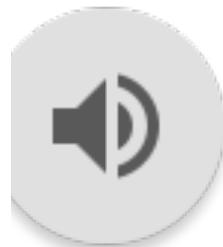
- Problème inhérent à la structure des LLMs
- Plus une phrase est répétée, plus la probabilité de continuer à générer cette phrase est élevée.
- Problème lié au langage qui favorise la répétition

II- Entraînement du Modèle : Problème de répétition



Exemples de résultats finaux, comparaison aux datasets

Résultat DadaGP



Extrait du dataset
(And Justice For All,
Metallica)



Résultat
1 GMP



Résultat
2 GMP



III- Résultats et Limitations

a) L'inutilité des nuances

```
p57:v95:d960:t0:b23 p52:v95:d960:t0:b23 p45:v95:d960:t0:b23 p57:v95:d960:t0:b21 p52:v95:d960:t0:b21 p45:v95:d960:t0:b21 p33:v95:d960:t960:b21  
p67:v95:d480:t0:b23 p62:v95:d6720:t0:b23 p55:v95:d480:t0:b23 p50:v95:d6720:t0:b23 p67:v95:d480:t0:b21 p62:v95:d480:t0:b21 p55:v95:d480:t0:b21  
p50:v95:d480:t0:b21 p31:v95:d480:t480:b21 p66:v95:d480:t0:b23 p62:v95:d0:t0:b23 p57:v95:d480:t0:b23 p50:v95:d0:t0:b23 p66:v95:d480:t0:b21  
p62:v95:d480:t0:b21 p57:v95:d480:t0:b21 p50:v95:d480:t0:b21 p30:v95:d480:t480:b21 p67:v95:d480:t0:b23 p62:v95:d0:t0:b23 p55:v95:d480:t0:b23  
p50:v95:d0:t0:b23 p67:v95:d480:t0:b21 p62:v95:d480:t0:b21 p55:v95:d480:t0:b21 p50:v95:d480:t0:b21 p31:v95:d480:t480:b21 p66:v95:d5280:t0:b23  
p62:v95:d0:t0:b23 p57:v95:d5520:t0:b23 p50:v95:d0:t0:b23 p66:v95:d960:t0:b21 p62:v95:d960:t0:b21 p57:v95:d960:t0:b21 p50:v95:d960:t0:b21  
p30:v95:d960:t960:b21 p66:v95:d0:t0:b23 p62:v95:d0:t0:b23 p57:v95:d0:t0:b23 p50:v95:d0:t0:b23 p66:v95:d4320:t0:b21 p62:v95:d4320:t0:b21  
p57:v95:d4320:t0:b21 p50:v95:d4320:t0:b21 p30:v95:d2880:t4320:b21 p57:v95:d0:t0:b23 p52:v95:d240:t0:b23 p45:v95:d240:t240:b23  
p57:v95:d240:t0:b23 p52:v95:d240:t0:b23 p45:v95:d240:t240:b23 p57:v95:d240:t0:b23 p52:v95:d240:t0:b23 p45:v95:d240:t480:b23 p55:v95:d240:t0:b23  
p45:v95:d240:t480:b23 p54:v95:d240:t0:b23 p45:v95:d240:t480:b23 p55:v95:d240:t0:b23 p45:v95:d240:t480:b23 p54:v95:d240:t0:b23  
p45:v95:d240:t480:b23 p55:v95:d240:t0:b23 p45:v95:d240:t480:b23 p57:v95:d720:t0:b23 p52:v95:d720:t0:b23 p45:v95:d720:t720:b23  
p57:v95:d240:t0:b23 p52:v95:d240:t0:b23 p45:v95:d240:t240:b23 p57:v95:d240:t0:b23 p52:v95:d240:t0:b23 p45:v95:d240:t480:b23 p55:v95:d240:t0:b23
```

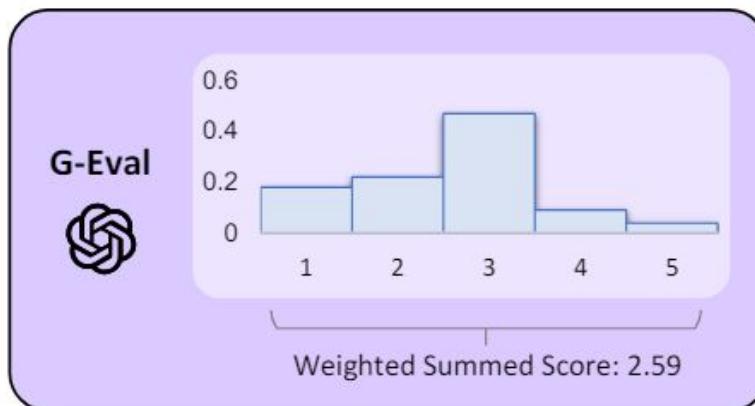
III- Résultats et Limitations

b) Limitation due à *Giant MIDI Piano*



III- Résultats et Limitations

c) L'absence de métriques objectives



Open LLM Leaderboard

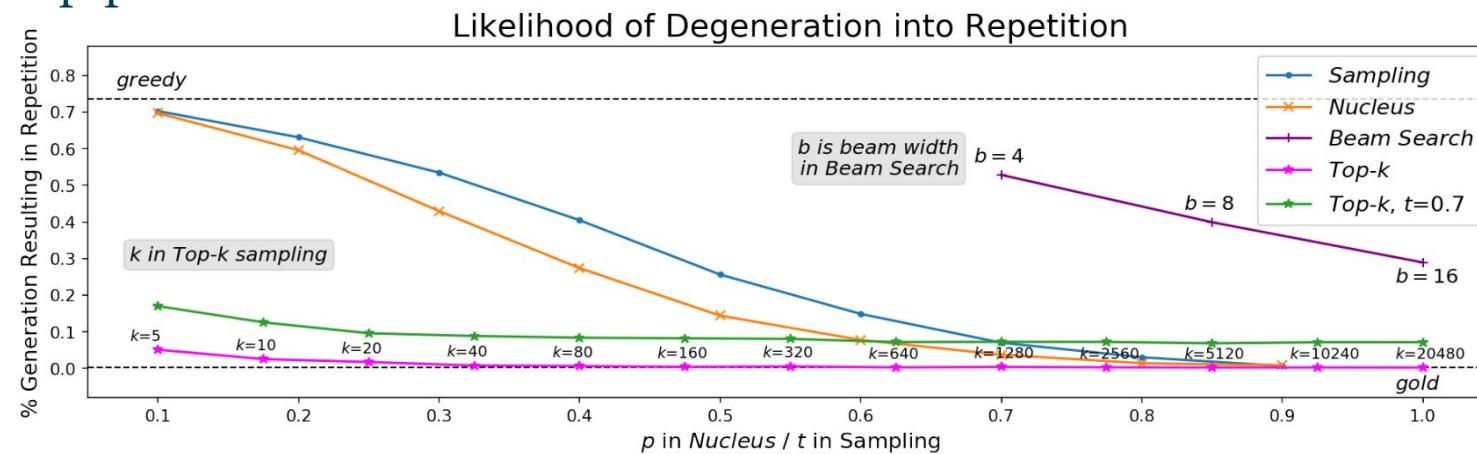
The screenshot shows a table from the Open LLM Leaderboard. The columns are "Revision", "Average", and "ARC (25-shot)". The rows list various LLM models. The row for "/falcon-40b" is highlighted with a red border.

	Revision	Average	ARC (25-shot)
/falcon-40b	main	60.4	61.9
s/llama-30b-supercot	main	59.8	58.5
65b	main	58.3	57.8
/GPT4-X-Alpasta-30b	main	57.9	56.7
us/Alpacino30b	main	57.4	57.1
GPT4-x-AlpacaDente2-30b	main	57.2	56.1
ke/Wizard-Vicuna-13B-Uncensored-HF	main	57	53.6
ke/dromedary-65b-lora-HF	main	57	57.8
30b	main	56.9	57.1
cess-ai-collective/wizard-mega-13b	main	55.7	52.5
ke/vicuna-13B-1.1-HF	main	53.7	47.4
lo/gpt4-x-alpaca	main	53.6	47.8

III- Résultats et Limitations

d) Le problème du *Pitch Bend* :

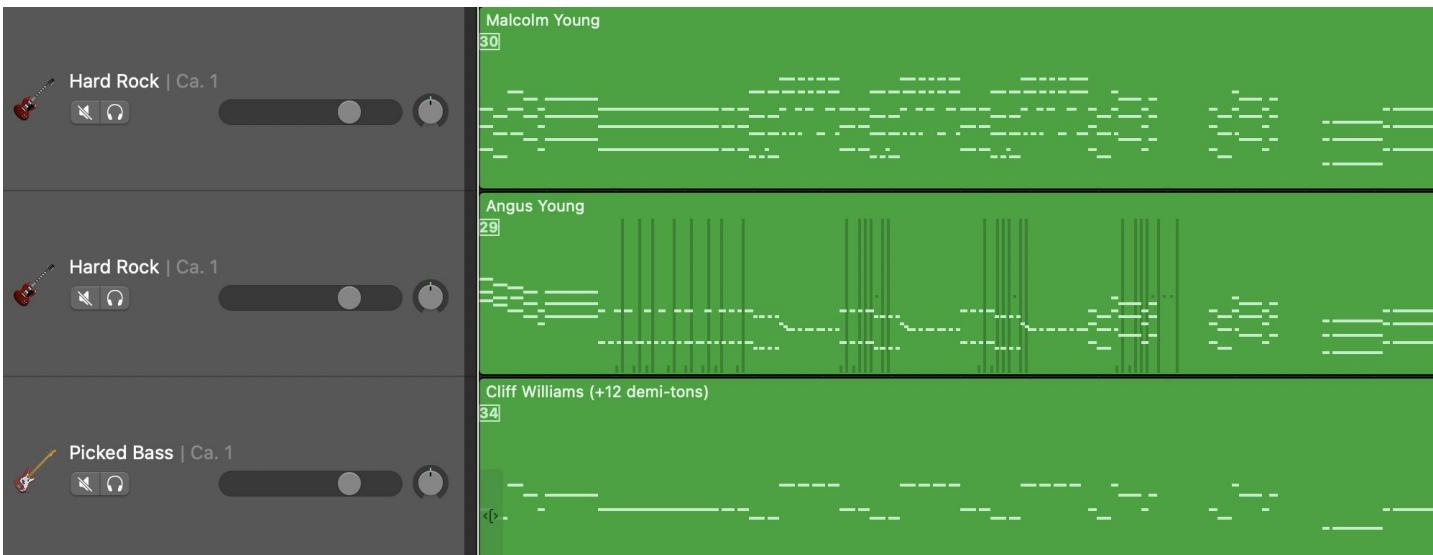
- La majorité des Pitch Bend du dataset sont quasiment sans effet :
 - zoom de notre quantification insuffisant
 - top-k vs top-p



- Réduire le dataset aux solos où le Pitch Bend est fortement utilisé : trop radical
- Beaucoup d'effet de Pitch Bend sont perdus dans la conversion GuitarProToMidi.

Conclusion :

- Une amélioration majeure est à attendre d'un changement radical de représentation :
 - *encodage temporel vs encodage piste par piste*
 - Il n'y a pas de bon solo de guitare sans un bon accompagnement : approche trop analytique.



```
drums:note:42
drums:note:38
wait:320
distorted0:note:s4:f3
distorted0:note:s5:f3
distorted0:note:s6:f1
distorted1:note:s2:f1
bass:note:s5:f1
wait:160
drums:note:42
wait:480
new_measure
distorted0:note:s4:f7
distorted0:note:s5:f7
distorted0:note:s6:f5
distorted1:note:s2:f1
drums:note:55
drums:note:42
drums:note:36
bass:note:s4:f0
wait:480
drums:note:42
wait:480
```

En format GuitarPro (ici en représentation textuelle), tous les instruments sont représentés sur une même piste. En format MIDI (ici en représentation *piano roll*), les pistes sont par essence séparées.

D'autres résultats de notre modèle

Rythme
syncopé

