sudo
Jurnal Teknik Informatika
https://jurnal.ilmubersama.com/index.php/sudo

Akal Imitasi

Pengembangan Sistem Generator MIDI Berbasis Web Menggunakan Arsitektur RWKV

Yusfiana[*], *Nahar Mardiyantoro, Muslim Hidayat*

Fakultas Teknik dan Komputer, Teknik Informatika, Universitas Sains Al-Qur'an, Wonosobo, Indonesia

INFORMASI ARTIKEL

Diterima Redaksi: 02 Juni 2026
Revisi Akhir: 22 Juni 2026
Diterbitkan *Online*: 30 Juni 2026

KATA KUNCI

Deep Learning
Generator Musik
Machine Learning
MIDI
RWKV

KORESPONDENSI (*)

Phone: +62 896-0368-8218
E-mail: yusfianaofficial@gmail.com

A B S T R A K

Pemanfaatan kecerdasan buatan (*Artificial Intelligence*/AI) dalam bidang kreatif, khususnya komposisi musik otomatis, terus berkembang seiring dengan kemajuan arsitektur deep learning. Berbagai pendekatan generatif telah dikembangkan, mulai dari model berbasis RNN, LSTM, GAN, VAE, hingga Transformer, namun sebagian besar masih memerlukan sumber daya komputasi yang besar untuk menghasilkan melodi yang koheren. Penelitian ini bertujuan mengembangkan sistem generator melodi MIDI berbasis web yang mampu menghasilkan komposisi musik secara otomatis dengan efisiensi tinggi. Sistem ini mengimplementasikan model *Receptance Weighted Key Value* (RWKV), sebuah arsitektur yang menggabungkan efisiensi komputasi *Recurrent Neural Network* (RNN) dengan kemampuan paralelisasi Transformer. Model dilatih menggunakan dataset yang bersumber dari 909 file MIDI mentah dengan ukuran konfigurasi *vocabulary* sebanyak 200 token unik, menghasilkan total keseluruhan 1.308.713 token sekuensial. Integrasi model ke platform web dilakukan menggunakan *framework* Gradio. Hasil pengujian teknis menunjukkan bahwa setelah melalui 5 epoch pelatihan pada dataset 1,3M token, model berhasil mencapai nilai *Cross-Entropy Loss* sebesar 0,8752 dengan tingkat *Perplexity* sebesar 2,40. Nilai ini jauh lebih rendah dibandingkan model LSTM pembanding yang memiliki *Perplexity* sekitar 18. Penggunaan arsitektur RWKV terbukti memberikan efisiensi sumber daya yang signifikan dengan penggunaan memori GPU sebesar 2,1 GB (hanya 14% dari kapasitas VRAM 15GB), yang memungkinkan proses generasi melodi secara real-time dengan latensi rendah.

PENDAHULUAN

Penelitian ini bertujuan untuk mengembangkan sistem generator melodi MIDI berbasis web yang memanfaatkan arsitektur *Receptance Weighted Key Value* (RWKV) sebagai solusi atas keterbatasan metode tradisional dan model sekuensial konvensional dalam menjaga koherensi melodi serta harmoni. Musik merupakan bahasa universal yang melibatkan pola temporal kompleks, sehingga pemodelannya membutuhkan arsitektur yang mampu menangkap ketergantungan jangka panjang secara efisien [1]. Meskipun teknik machine learning seperti LSTM telah banyak digunakan untuk mempelajari pola musikal yang kompleks [2]. Penggunaan arsitektur RWKV diharapkan mampu memberikan efisiensi komputasi yang lebih baik dalam memproses data sekuensial MIDI tanpa mengorbankan kualitas komposisi yang dihasilkan.

Melalui proses ekstraksi nada, pembentukan token unik, hingga implementasi pada platform berbasis web, sistem ini dirancang untuk menghasilkan melodi baru yang variatif namun tetap terdengar alami dan terstruktur. Dengan demikian, proyek ini berkontribusi pada pengembangan teknologi generatif yang menggabungkan keunggulan *deep learning modern* dengan aksesibilitas perangkat lunak berbasis peramban bagi para kreator musik.

Model Transformer, yang awalnya diperkenalkan untuk penerjemahan bahasa, telah membuktikan keunggulannya dalam memproses data teks, audio, hingga gambar, serta sangat efektif dalam mentransfer pengetahuan melalui teknik pre-training seperti autoregressive atau masked language modeling. Namun, penggunaan model ini sering terkendala oleh

https://doi.org/10.56211/sudo.v5i2.1777

biaya komputasi dan penggunaan memori yang sangat tinggi akibat mekanisme *self-attention* yang memiliki kompleksitas kuadratis $O(N^2)$, sehingga membatasi panjang konteks dan menghambat penangkapan ketergantungan jangka panjang (*long-term dependencies*). Permasalahan ini semakin nyata ketika diterapkan pada generasi musik simbolik yang dapat mencapai lebih dari 10.000 token [3,4].

Meskipun beberapa penelitian mencoba mengoptimasi kompleksitas tersebut menjadi sparse atau menggunakan *locality-sensitive hashing*, metode-metode tersebut umumnya belum mampu mempercepat proses *autoregressive inference*. Sebagai solusi, diperkenalkan model *linear* Transformer yang memanfaatkan formulasi berbasis kernel dan sifat asosiatif perkalian matriks untuk mencapai skalabilitas linear terhadap panjang konteks [4]. Inovasi ini tidak hanya mengurangi beban memori secara signifikan, tetapi juga mengungkap keterkaitan antara Transformer dan RNN, sehingga memungkinkan proses inference berjalan hingga ribuan kali lebih cepat dengan performa yang tetap setara dengan model Transformer standar.

Deep learning telah membawa kemajuan besar dalam kecerdasan buatan, terutama untuk memproses data sekuensial kompleks seperti bahasa alami, analisis time-series, hingga format gambar dan graf. Untuk menjembatani celah antara efisiensi RNN dan skalabilitas Transformer, diperkenalkan arsitektur RWKV (*Receptance Weighted Key Value*) yang menggabungkan efisiensi linear milik RNN dengan kemampuan skalabilitas serta pelatihan paralel khas Transformer. Dengan memformulasikan ulang mekanisme *attention* menggunakan varian *linear attention* yang lebih efektif, RWKV mampu memproses miliaran parameter dengan biaya komputasi dan penggunaan memori yang jauh lebih rendah tanpa mengorbankan performa [5]. Pengembangan lanjutan arsitektur ini, yakni RWKV-5 (Eagle) dan RWKV-6 (Finch), semakin meningkatkan ekspresivitas model melalui mekanisme *multi-headed matrix-valued states* dan *dynamic recurrence* [6].

Mengingat output dari penelitian ini bersifat generatif (komposisi musik), maka keberhasilan model tidak diukur menggunakan metrik klasifikasi tradisional seperti Accuracy atau F1-Score. Sebaliknya, penelitian ini menerapkan standar evaluasi *probabilistic modeling* melalui nilai *Cross-Entropy Loss*, RMSE, MAPE, dan *Perplexity*. Metrik ini dipilih karena mampu merepresentasikan seberapa akurat arsitektur RWKV dalam memprediksi distribusi nada MIDI berikutnya berdasarkan pola yang dipelajari dari dataset [7].

TINJAUAN PUSTAKA

Deep Learning dalam Musik

Deep learning dalam musik melibatkan pelatihan jaringan saraf tiruan, khususnya model sekuensial seperti *Long Short-Term Memory* (LSTM), untuk mempelajari pola dan ketergantungan temporal dari data musik seperti berkas MIDI. [2] membuktikan bahwa model *deep learning* berbasis LSTM mampu mempelajari pola jazz MIDI dan menghasilkan komposisi musik baru yang serupa dengan *dataset* aslinya. Dalam penelitian tersebut, model dilatih menggunakan data MIDI genre jazz dengan instrumen tunggal, menghasilkan kelanjutan melodi yang koheren melalui prediksi token berikutnya (*next-token prediction*).

Studi lebih lanjut oleh [8] mengembangkan *pipeline* generasi musik otomatis berbasis LSTM menggunakan representasi simbolik MIDI dari musik piano klasik. Model tiga-layer LSTM tersebut dilatih selama 100 epoch dan dievaluasi menggunakan metrik *loss*, *accuracy*, *top-3 accuracy*, dan *perplexity*. Hasilnya menunjukkan *validation loss* sebesar 2,93 dan *perplexity* sekitar 18, yang mengindikasikan kemampuan prediksi sekuens yang memadai.

Namun demikian, meskipun LSTM sangat efektif dalam menangani urutan data, model ini sering kali terhambat oleh keterbatasan dalam memproses dependensi jangka sangat panjang serta sulit untuk diparalelisasi saat pelatihan. Sebagai solusi, penelitian ini mengimplementasikan arsitektur RWKV yang menggabungkan efisiensi komputasi linier $O(T)$ milik model rekurens (RNN) dengan kemampuan skalabilitas dan pelatihan paralel layaknya model Transformer.

Representasi MIDI

MIDI (*Musical Instrument Digital Interface*) merupakan format data musik simbolik dasar yang digunakan secara luas dalam pembuatan musik berbasis deep learning. Hal ini dikarenakan efisiensinya dalam menerjemahkan informasi musik menjadi urutan peristiwa (*events*) seperti nada (*pitch*) dan durasi. Dalam sistem ini, data MIDI diekstraksi menggunakan pustaka pretty_midi dan melalui proses tokenisasi menjadi urutan numerik agar dapat diproses oleh arsitektur RWKV sebagai tugas prediksi token berikutnya. Representasi simbolik ini dipilih karena kemampuannya menangkap atribut tingkat nada yang krusial namun tetap hemat dalam penggunaan sumber daya komputasi [1].

[9] mengusulkan pendekatan MIDI2vec, yaitu representasi file MIDI sebagai vektor melalui teknik *graph embedding*. Pendekatan ini merepresentasikan data MIDI sebagai graf yang mencakup informasi tempo, *time signature*, program, dan nada, kemudian menjalankan algoritma node2vec untuk menghasilkan *embedding* melalui random walks pada graf tersebut. Hasil *embedding* terbukti berhasil digunakan untuk memprediksi genre musik dan metadata lainnya seperti komposer, instrumen, dan gerakan musik dengan akurasi yang baik.

Untuk penanganan sekuens MIDI yang panjang, [10] mengusulkan *Multitrack Music* Transformer (MMT) dengan representasi musik multi-instrumen baru yang memungkinkan penggunaan sekuens yang lebih pendek namun tetap mencakup beragam instrumen. MMT mencapai performa setara dengan sistem *state-of-the-art* namun dengan *speedup* substansial dan pengurangan memori yang signifikan, sehingga cocok untuk aplikasi improvisasi *real-time*.

Arsitektur RWKV (Receptance Weighted Key Value

Implementasi arsitektur RWKV dalam penelitian ini didasarkan pada kemampuannya menggabungkan efisiensi komputasi linier *O(L)* milik *Recurrent Neural Networks* (RNN) dengan skalabilitas pelatihan paralel khas Transformer. Berbeda dengan model Transformer tradisional yang memiliki kompleksitas kuadratis $O(L^2)$ sehingga boros memori pada urutan data panjang, RWKV memungkinkan pemrosesan sekuensial yang lebih cepat dengan penggunaan sumber daya yang jauh lebih rendah [5].

[6] kemudian memperkenalkan RWKV-5 (Eagle) dan RWKV-6 (Finch), yang merupakan penyempurnaan atas arsitektur RWKV-4 dengan penambahan mekanisme *multi-headed matrix-valued states* dan *dynamic recurrence* untuk meningkatkan ekspresivitas model sambil mempertahankan efisiensi inferensi khas RNN. Model Eagle mencakup empat varian dengan parameter antara 0,46 hingga 7,5 miliar, sementara model Finch tersedia dalam ukuran 1,6 dan 3,1 miliar parameter.

[11] memperluas aplikasi RWKV ke domain visi melalui Vision-RWKV (VRWKV), sebuah model yang memodifikasi arsitektur RWKV dengan *bidirectional spatial processing* dan *stability layers* untuk tugas persepsi visual. VRWKV terbukti melampaui performa ViT dalam klasifikasi gambar dengan kecepatan yang jauh lebih tinggi dan penggunaan memori yang lebih rendah.

Pendekatan* Transformer *untuk Generasi Musik Simbolik

Perkembangan model Transformer dalam generasi musik simbolik telah membuka peluang baru dalam komposisi musik otomatis. [10] dengan MMT-nya membuktikan bahwa dengan representasi yang tepat, model Transformer dapat digunakan untuk generasi musik multi-instrumen secara efisien. Model ini menggunakan *decoder-only* Transformer dengan input dan output multi-dimensi untuk mengurangi kompleksitas memori.

Pendekatan berbeda diajukan oleh [3] melalui Museformer, sebuah Transformer dengan mekanisme *attention fine-grained* dan *coarse-grained* yang dirancang khusus untuk generasi musik. Dengan *fine-grained attention*, sebuah token dapat langsung memperhatikan semua token dari bar-bar yang paling relevan secara struktural (misalnya bar ke-1, 2, 4, dan 8 sebelumnya). Sementara itu, *coarse-grained attention* hanya memperhatikan ringkasan dari bar-bar lainnya untuk mengurangi biaya komputasi. Hasilnya, Museformer mampu memodelkan sekuens musik 3x lebih panjang dibandingkan full-attention counterpart-nya, sambil tetap menghasilkan musik berkualitas tinggi dengan struktur yang lebih baik.

[12] menerapkan *Large Language Models* (LLMs) pada *pre-training* musik simbolik dan menemukan bahwa LLM lebih kompatibel dengan notasi ABC dibandingkan MIDI karena keselarasannya dengan desain LLM. Mereka mengusulkan *Synchronized Multi-Track* ABC Notation (SMT-ABC Notation) untuk menjaga koherensi lintas beberapa trek musik. Hasil eksperimen menunjukkan bahwa pendekatan mereka mengungguli GPT-4 sebesar 17% dalam metrik repetisi musik dan melampaui model ChatMusician sebesar 6% pada test set single-track.

Linear* Transformer *dan Efisiensi Komputasi

[4] memperkenalkan Linear Transformer yang mengekspresikan *self-attention* sebagai *dot-product* linier dari *kernel feature maps*, memanfaatkan sifat asosiatif perkalian matriks untuk mengurangi kompleksitas dari $O(N^2)$ menjadi *O(N)*. Formulasi ini memungkinkan implementasi iteratif yang secara dramatis mempercepat *autoregressive* Transformer dan mengungkap hubungannya dengan *Recurrent Neural Networks*. Linear Transformer mencapai performa serupa dengan vanilla Transformer dan hingga 4000x lebih cepat pada prediksi *autoregressive* untuk sekuens yang sangat panjang.

Survei Deep Learning untuk Generasi Musik Simbolik

[1] menyajikan survei komprehensif tentang *deep learning* untuk generasi musik simbolik, mencakup representasi data, algoritma, evaluasi, dan tantangan yang ada. Survei ini mengkategorikan berbagai pendekatan termasuk RNN, LSTM, Transformer, GAN, VAE, dan model hibrida. Studi ini mengidentifikasi bahwa representasi data MIDI merupakan salah satu faktor kunci dalam keberhasilan sistem generasi musik, dan bahwa pemilihan metrik evaluasi yang tepat sangat penting mengingat sifat subjektif dari kualitas musik.

Teori Metrik Evaluasi (Cross-Entropy & Perplexity)

Untuk mengukur akurasi prediksi model, digunakan *Cross-Entropy Loss* yang secara matematis terbukti memiliki jaminan batas konsistensi-H (*H-consistency bounds*) dalam memperkecil kesalahan aktual. [7] mempresentasikan analisis teoretis dari keluarga luas fungsi kerugian bernama *comp-sum losses*, yang mencakup *cross-entropy* atau *logistic loss*, *generalized cross-entropy*, dan *mean absolute error*. Mereka memberikan *H-consistency bounds* pertama untuk fungsi-fungsi kerugian ini. Selain itu, tingkat keyakinan model dievaluasi menggunakan *Perplexity*, di mana skor yang lebih rendah menandakan bahwa model lebih yakin dan akurat dalam menghasilkan token musik yang koheren. [8] mengkonfirmasi bahwa *perplexity* merupakan indikator yang efektif untuk model generasi musik berbasis LSTM. Perbandingan nilai *perplexity* antara model RWKV (2,40) dan LSTM (18) dalam penelitian ini memberikan bukti kuantitatif tentang keunggulan arsitektur RWKV dalam konteks generasi melodi MIDI.

Penelitian Terdahulu

Tabel berikut merangkum penelitian terdahulu yang relevan sebagai landasan penelitian ini.

Tabel 1. Penelitian Terdahulu

No	Penulis & Tahun	Topik Penelitian	Hasil Utama
1	[13]Muhammad Djamaluddin et al. (2026)	Pengembangan Model Deep Learning Long Short-Term Memory untuk Generasi Musik Berbasis Data MIDI	Membuktikan bahwa arsitektur LSTM mampu mempelajari pola sekuens musik dengan baik sebagai model prediksi sekuensial berbasis RNN.
2	[12] Qu et al. (2024)	Penerapan LLMs untuk generasi musik simbolik dengan notasi ABC	LLM lebih kompatibel dengan notasi ABC dibanding MIDI; model mengungguli GPT-4 sebesar 17% pada metrik repetisi musik dan melampaui ChatMusician 6%
3	[4]Katharopoulos et al. (2020)	Optimasi Transformer melalui formulasi berbasis kernel (Linear Transformer)	Inference berjalan hingga 4000x lebih cepat dengan performa setara model Transformer standar melalui kompleksitas O(N)
4	[5]Peng et al. (2023)	Memperkenalkan arsitektur RWKV sebagai alternatif Transformer untuk pemrosesan sekuensial	RWKV mampu memproses miliaran parameter dengan biaya komputasi rendah; model terbesar mencapai 14 miliar parameter
5	[6]Peng et al. (2024)	Pengembangan RWKV-5 (Eagle) dan RWKV-6 (Finch) dengan mekanisme multi-headed matrix-valued states	Mencapai performa kompetitif pada berbagai benchmark dengan model hingga 7,5 miliar parameter
6	[11]Duan et al. (2025)	Vision-RWKV (VRWKV): aplikasi RWKV pada tugas persepsi visual	VRWKV melampaui ViT dalam klasifikasi gambar dengan kecepatan lebih tinggi dan memori lebih rendah
7	[1]Ji et al. (2023)	Survei komprehensif deep learning untuk generasi musik simbolik	Memetakan state-of-the-art metode generasi musik; mengidentifikasi MIDI sebagai representasi dominan

 https://doi.org/10.56211/sudo.v5i2.1777

8	[2]Yadav et al. (2022)	Model deep learning ringan untuk generasi musik jazz berbasis MIDI dengan LSTM	Model LSTM berhasil mempelajari pola jazz MIDI dan menghasilkan komposisi baru; validasi pendekatan next-token prediction
9	[8]Nagoshi (2025)	Generasi musik otomatis berbasis LSTM menggunakan representasi simbolik MIDI piano klasik	Validation loss 2,93; *perplexity* ~18; top-3 accuracy 0,53; dokumentasi metodologi sistematis
10	[10]Dong et al.(2023) - MMT	Multitrack Music Transformer (MMT) dengan representasi musik multi-instrumen baru	Performa sebanding dengan sistem SOTA dengan speedup substansial dan pengurangan memori; cocok untuk improvisasi real-time
11	[3]Yu et al. (2022) - Museformer	Transformer dengan fine- dan coarse-grained attention untuk generasi musik	Mampu memodelkan sekuens musik 3x lebih panjang; menghasilkan musik berkualitas tinggi dengan struktur lebih baik
12	[9]Lisena et al. (2022)	MIDI2vec: Representasi file MIDI sebagai vektor menggunakan graph embedding	Embedding MIDI berbasis graph berhasil memprediksi genre, komposer, dan instrumen dengan akurasi baik
13	[7]Mao et al. (2023)	Analisis teoretis *Cross-Entropy Loss* dan keluarga comp-sum losses	Memberikan H-consistency bounds pertama untuk cross-entropy; membuktikan meminimalkan cross-entropy memperkecil kesalahan klasifikasi aktual

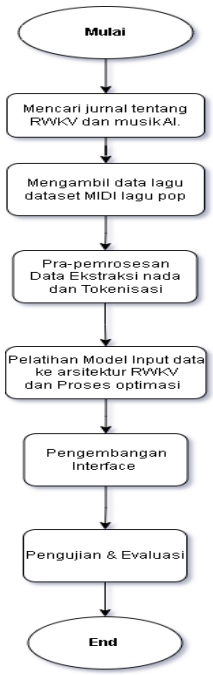
Kerangka Pemikiran

Penelitian ini menjembatani celah antara penggunaan RNN yang hemat memori namun sulit diparalelisasi, dengan Transformer yang unggul dalam ketergantungan jarak jauh namun boros sumber daya. Dengan mengimplementasikan RWKV, sistem diharapkan mampu melakukan prediksi nada berikutnya (*next-token prediction*) secara efisien. Hubungan antar variabel dalam penelitian ini melibatkan input berupa Seed (nada pemicu), Length, dan Konfigurasi tangga nada (kombinasi *Root Note* dan *Scale/Mode* seperti Dorian, Phrygian, Lydian, dll.), yang diproses melalui 8 layer blok RWKV untuk menghasilkan output berupa melodi MIDI yang koheren namun tetap variatif melalui teknik sampling probabilitas.

METODOLOGI

Alur Penelitian

Penelitian ini dilakukan melalui beberapa tahapan sistematis bisa dilihat pada Gambar 1 dibawah, dimulai dari pengumpulan dataset MIDI, pra-pemrosesan data, pelatihan model RWKV, hingga tahap evaluasi menggunakan antarmuka interaktif. Pendekatan ini konsisten dengan metodologi yang direkomendasikan oleh [1] untuk sistem generasi musik simbolik berbasis deep learning.



Gambar 1. *Flowchart* Alur Penelitian

Pengumpulan dan Pra-pemrosesan Data

Data yang digunakan dalam penelitian ini adalah dataset musik digital dalam format MIDI (*Musical Instrument Digital Interface*) yang telah dilabeli antara akord minor dan major. Tahap pra-pemrosesan dilakukan dengan mengekstraksi informasi *pitch* dan durasi dari setiap file MIDI menggunakan pustaka *pretty_midi*. Total data yang dikumpulkan mencakup 909 file MIDI mentah yang kemudian dikonversi menjadi urutan token numerik (tokenisasi). Proses tokenisasi ini menghasilkan 200 token unik pada kamus data (*vocabulary size*) dengan total keseluruhan mencapai 1.308.713 token sekuensial. Setiap token merepresentasikan nada unik dalam rentang standar MIDI (0-127).

Proses tokenisasi tersebut menghasilkan urutan sekuensial angka yang merepresentasikan tinggi nada (*pitch*) dan durasi waktu, yang kemudian siap diumpankan ke dalam model untuk proses pembelajaran. Untuk mengoptimalkan kualitas harmonisasi melodi yang dihasilkan dalam keterbatasan pelatihan 5 epoch, sistem menerapkan modul pembatas aturan (*rule-based constraints*) pada tahap inferensi menggunakan fungsi *snap_to_scale* berbasis kamus *data preset musical* yang otomatis mengatur *Root Note* (nada dasar) dan *Scale/Mode* untuk memfilter serta memprioritaskan probabilitas kemunculan token nada tertentu agar terhindar dari nada sumbang. Daftar pustaka tangga nada dan preset suasana yang diimplementasikan pada antarmuka sistem dapat dilihat pada Tabel 2:

Tabel 2. Pustaka Konfigurasi Nada Dasar dan Skala (*Scale/Mode*) pada Sistem

No	Nama Pilihan pada Interface	Root Note (Nada Dasar)	Jenis Tangga Nada (Scale Mode)
1	Cerah/Pop	C	Major Ionian
2	Sedih/Ballad	A	Minor (Aeonian)
3	Jazz/Groove	D	Dorian
4	Orientel	E	Phrygian
5	Epik	F	Lydian
6	Blues	G	Blues
7	Folk/Simple	C	Major Pentatonic

Arsitektur Model dan Konfigurasi Hyperparameter

Penelitian ini mengimplementasikan arsitektur RWKV yang menggabungkan struktur rekurens dengan efisiensi Transformer. Model yang dirancang memiliki spesifikasi struktur berupa 8 blok layer RWKV dengan embedding size

 https://doi.org/10.56211/sudo.v5i2.1777

sebesar 256 dimensi. Penggunaan Linear Attention diterapkan untuk memastikan kompleksitas waktu yang linier *O(T)* terhadap panjang sekuens music, untuk proses pelatihan dan proses inferensi real-time pada sistem, konfigurasi *hyperparameter* diatur secara spesifik pada Tabel 3:

Tabel 3. Konfigurasi *Hyperparameter*

Parameter	Nilai/Konfigurasi
Batch Size	1
Learning Rate	$5x10^{-4}$
Token per Bar	1.000
Mekanisme Inferensi (Sampling)	Kombinasi Top-K ($k = 40$) dan Top-P *Nucleus Sampling* ($p = 0{,}95$)

Model dilatih menggunakan optimasi Adam dengan *learning rate* sebesar 5×10^{-4} untuk meminimalkan nilai *Cross-Entropy Loss* selama proses prediksi nada berikutnya (next-token prediction). Pemilihan optimasi Adam dan *learning rate* ini konsisten dengan praktik terbaik dalam pelatihan model sekuensial untuk generasi musik [8]. Nilai *Perplexity* juga dihitung untuk mengukur seberapa baik model dapat memprediksi atribut musik berikutnya seperti nada (*pitch*), durasi, atau ketukan (*velocity*).

$$wkv_t = \frac{\sum_{i=1}^{t-1} e^{-(t-1-i)w+k_i} v_i + e^{u+k_t} v_t}{\sum_{i=1}^{t-1} e^{-(t-1-i)w+k_i} + e^{u+k_t}}$$

Secara matematis, mekanisme inti RWKV direpresentasikan melalui persamaan *weighted average* yang interaksinya dikontrol secara eksponensial. Representasi informasi musik dari masa lalu (v_i) digabungkan dengan informasi pada waktu saat ini (v_t) menggunakan bobot dari vektor Key (k_i dan k_t). Melalui penggunaan fungsi eksponensial berbasis peluruhan waktu (*Time Decay* atau *w*), bobot informasi masa lalu akan mengecil secara berkala seiring bertambahnya jarak waktu ($t - 1 - i$). Karakteristik matematis inilah yang memungkinkan akumulasi riwayat melodi disimpan ke dalam satu variabel penampung (*state memory*), sehingga kompleksitas kuadratik Transformer dapat dipangkas menjadi kompleksitas linier *O(T)* [5].

Cross-Entropy Loss

Cross-Entropy Loss adalah fungsi kerugian (*loss function*) yang sangat populer dalam pembelajaran mendalam, khususnya untuk masalah klasifikasi, karena mengukur seberapa baik probabilitas prediksi model (yang biasanya dihasilkan melalui fungsi aktivasi *softmax*) selaras dengan label target yang sebenarnya (*ground truth*). [7] membuktikan secara matematis bahwa *cross-entropy* termasuk dalam keluarga besar fungsi kerugian bernama *comp-sum losses*, dan memberikan *H-consistency bounds* sebagai jaminan non-asimtotik yang membuktikan bahwa *meminimalkan cross-entropy* secara efektif dan ketat akan memperkecil kesalahan klasifikasi aktual.

Dalam penelitian ini, *Cross-Entropy Loss* pada generator MIDI AI berfungsi untuk menjaga struktur lagu agar tetap harmonis dari awal hingga akhir. Arsitektur RWKV memiliki mekanisme attention linier yang memungkinkannya mengingat struktur atau motif musik yang sudah lewat (konteks panjang) dengan jauh lebih efisien daripada RNN tradisional seperti LSTM, sehingga membantu menjaga nilai RMSE dan MAPE evaluasi tetap optimal.

Perplexity

Perplexity adalah metrik umum yang digunakan untuk mengevaluasi kinerja model sekuensial, termasuk model pembuat musik. Metrik ini mengukur seberapa baik model memprediksi elemen berikutnya dalam suatu urutan, di mana nilai *perplexity* yang lebih rendah menandakan prediksi yang lebih akurat dan pasti. [8] mengkonfirmasi bahwa *perplexity* merupakan indikator yang efektif untuk model generasi musik berbasis LSTM, dan nilai *perplexity* yang diperoleh dalam penelitiannya (~18) memberikan *baseline* perbandingan yang jelas terhadap hasil penelitian ini (2,40), menunjukkan keunggulan signifikan arsitektur RWKV.

Perancangan Interface

Untuk menguji kemampuan model secara *real-time*, dikembangkan sebuah sistem antarmuka berbasis web menggunakan *framework* Gradio. Platform ini dirancang agar ramah pengguna (*user-friendly*) dengan meniadakan input nada awal (*seed*). Proses pembangkitan melodi sepenuhnya dikendalikan oleh dua parameter utama yang dimasukkan oleh pengguna

melalui antarmuka, yaitu jumlah birama musik (*Bar Options*) dengan rentang pilihan 2 hingga 32 bar, serta pemilihan Preset Suasana Cepat (*Mood Presets*). Hasil akhir dari proses inferensi ini berupa sekuens angka yang dikonversi kembali menjadi file berekstensi .mid (MIDI) yang dapat langsung diputar atau diunduh oleh pengguna melalui halaman web

Dalam proses pembangkitan melodi, sistem menggunakan teknik *probability sampling*. Ketika pengguna menentukan pilihan preset tertentu pada antarmuka, fungsi apply_mood pada kodingan sistem akan langsung mengunci rumus tangga nada yang sesuai. Algoritma inferensi kemudian melakukan pemotongan probabilitas (*pruning*) atau memberikan bobot matematika yang lebih tinggi (*bias adjustment*) pada token MIDI (rentang 0–127) yang hanya patuh pada rumus interval tangga nada yang dipilih melalui fungsi snap_to_scale. Melalui mekanisme pembatasan berbasis aturan (*rule-based constraints*) ini, model RWKV dipandu untuk tetap menghasilkan progresi nada yang harmonis dan terstruktur sesuai dengan teori musik formal, sehingga mampu menekan potensi timbulnya anomali nada fales meskipun model dilatih dalam lingkungan komputasi dengan sumber daya terbatas.

Lingkungan Eksperimen

Eksperimen dilakukan pada lingkungan dengan spesifikasi *hardware* dan *software* sebagai berikut:

Tabel 4. Spesifikasi Lingkungan Eksperimen

Komponen	**Spesifikasi**
Prosesor	Intel® Xeon® (Google Colaboratory)
RAM	112.7 GB (Google Colaboratory)
GPU	NVIDIA Tesla T4 16GB VRAM
Platform	Google Colaboratory (Free Tier)
Operating System	Linux Ubuntu (Cloud Environment)
Software	Python 3.10.11, PyTorch, dan Gradio
File MIDI mentah	909 File
Total token hasil tokenisasi	1.308.713
Token Unik	200

HASIL DAN PEMBAHASAN

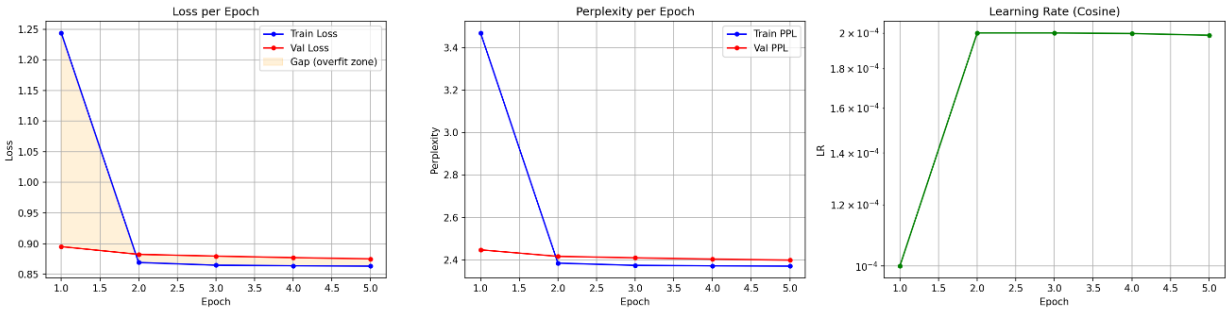
Evaluasi Konvergensi Model

Berdasarkan observasi pembelajaran, nilai *loss* terbaik 0,8752 yang dicapai pada epoch ke-5 telah menunjukkan titik jenuh awal (*initial convergence*), di mana model sudah mampu merepresentasikan pola sekuensial MIDI dengan akurasi yang memadai untuk kebutuhan generator melodi. Evaluasi kuantitatif terhadap model generator melodi dilakukan menggunakan metode *out-of-sample testing* dengan melibatkan 78 sampel file MIDI sebagai data uji terpisah. Ukuran sampel ini ditentukan berdasarkan proporsi data uji sebesar 8-10% dari total populasi dataset (1.308.713 token).

Pencapaian *Cross-Entropy Loss* sebesar 0,8752 dan *Perplexity* sebesar 2,40 setelah 5 epoch pelatihan menunjukkan efisiensi luar biasa dari arsitektur linier RWKV. Sebagai perbandingan, model LSTM dalam penelitian terdahulu memerlukan 100 epoch penuh hanya untuk mencapai tingkat *perplexity* di kisaran 18. Penggunaan alokasi memori GPU Tesla T4 tercatat stabil di angka 2,1 GB (sekitar 14% dari kapasitas total VRAM 15GB yang tersedia), membuktikan bahwa model ini sangat ringan dan sangat memadai jika diintegrasikan langsung pada server web berbiaya rendah untuk keperluan generasi melodi secara *real-time*.

Dinamika proses pelatihan secara keseluruhan dapat diamati pada Gambar 2, yang menampilkan kurva *Train Loss*, *Val Loss*, *Perplexity*, dan *Learning Rate* selama 5 epoch. Kurva menunjukkan konvergensi yang stabil tanpa indikasi *overfitting*, dengan gap antara *train* dan *val loss* yang semakin mengecil seiring bertambahnya epoch.

 https://doi.org/10.56211/sudo.v5i2.1777



Gambar 2. RWKV-Train vs Val Monitor (1.3 Token, T4)

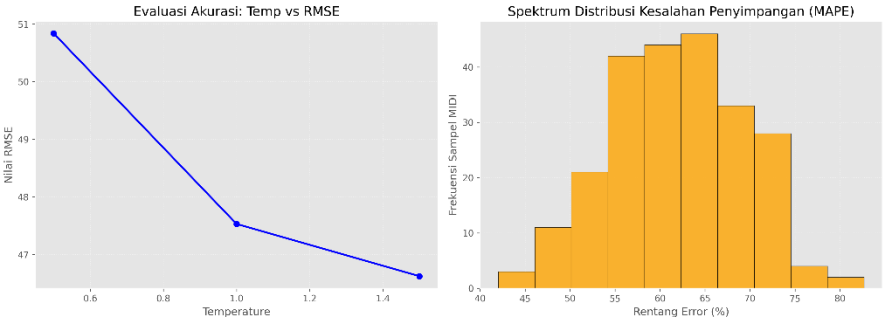
Pada Gambar 2 kurva pelatihan model RWKV selama 5 epoch pada dataset 1,3M token menggunakan GPU Tesla T4. (Kiri) *Loss per Epoch* menunjukkan Train Loss dan Val Loss yang keduanya konvergen secara stabil mulai epoch ke-2, dengan gap yang mengecil dan tidak menunjukkan indikasi *overfitting*. (Tengah) *Perplexity per Epoch* mengonfirmasi penurunan konsisten hingga mencapai Val PPL sebesar 2,40 pada epoch ke-5. (Kanan) Jadwal *Learning Rate* berbasis Cosine Annealing yang memastikan kestabilan proses optimasi.

Evaluasi Akurasi Hasil Generasi

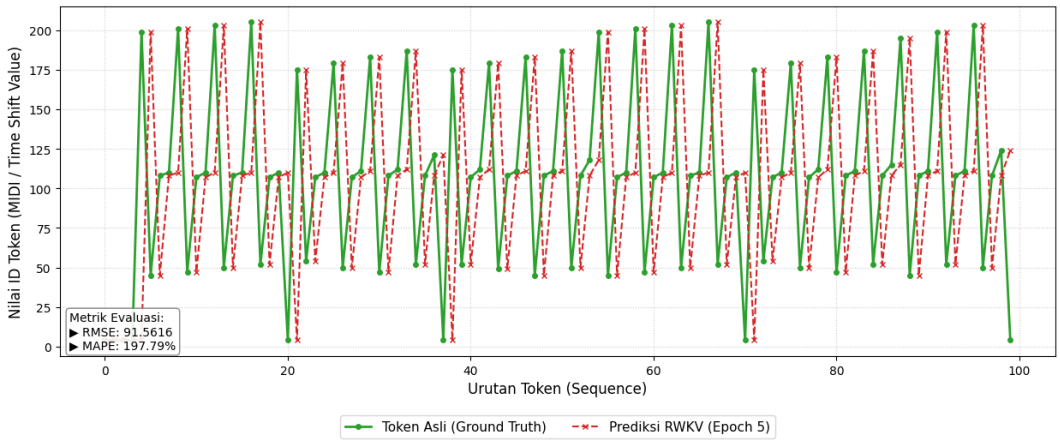
Berdasarkan hasil pengujian luar sampel (*out-of-sample testing*) dengan 78 sampel MIDI yang dapat dilihat pada Gambar 3, parameter Temperature terbukti memiliki pengaruh signifikan terhadap akurasi rekonstruksi melodi maupun tingkat kepatuhan tangga nada. Peningkatan Temperature dari 0,5 ke 1,5 justru menyebabkan nilai RMSE menurun dari 50,8382 menjadi 46,6238, dan nilai MAPE turut menurun dari 65,71% menjadi 59,22%. Sementara itu, tingkat kepatuhan nada (*compliance*) terhadap tangga nada meningkat dari 64,87% menjadi 71,41%, dengan rasio nada fales yang berkurang dari 35,13% menjadi 28,59%. Temuan ini mengindikasikan bahwa Temperature yang lebih tinggi tidak serta-merta mengorbankan koherensi harmonik, melainkan memberikan fleksibilitas eksplorasi probabilistik yang justru membantu model menghindari nada-nada di luar tangga nada. Meskipun demikian, tingkat leap error tetap tinggi di kisaran 74–76% pada seluruh variasi Temperature, menunjukkan bahwa anomali lompatan melodi merupakan tantangan tersendiri yang tidak dipengaruhi secara signifikan oleh parameter Temperature.

Temuan ini konsisten dengan prinsip sampling probabilistik dalam generasi musik yang diidentifikasi oleh [1] di mana terdapat *trade-off* antara kreativitas (variasi) dan koherensi (akurasi). Parameter Temperature berfungsi sebagai pengendali *trade-off* tersebut, memungkinkan pengguna untuk menyesuaikan karakteristik output sesuai kebutuhan kreatif mereka.

Visualisasi perbandingan antara token asli dan prediksi model pada 100 token pertama dapat dilihat pada Gambar 4. Pola prediksi secara umum mengikuti fluktuasi sekuensial token asli, meskipun terdapat selisih nilai pada token-token yang merepresentasikan Time Shift Value dengan skala lebih besar, yang menjadi penyebab utama tingginya nilai RMSE dan MAPE secara keseluruhan.



Gambar 3. Grafik Evaluasi Akurasi Teknis Generasi



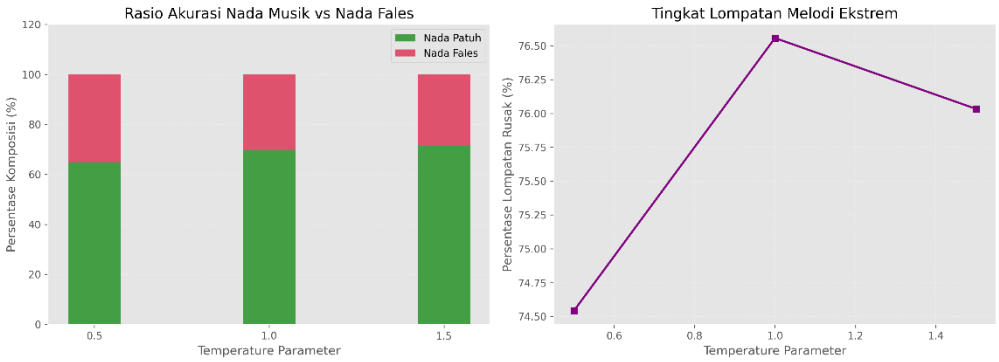
Gambar 4. Grafik Perbandingan Token Asli vs Prediksi RWKV

Pada Gambar 3 perbandingan token asli (*ground truth*) dengan prediksi model RWKV pada 100 token pertama (Epoch 5). Pola fluktuasi prediksi mengikuti struktur sekuensial token asli dengan RMSE sebesar 91,5616 dan MAPE sebesar 197,79%, yang mencerminkan karakteristik prediksi pada ruang token campuran antara nilai MIDI pitch (0–127) dan Time Shift Value yang memiliki skala berbeda.

Evaluasi Tingkat Nada Fales & Anomali Melodi

Tingkat nada fales yang terukur pada pengujian *out-of-sample* berkisar antara 28,59% hingga 35,13% tergantung parameter Temperature, namun perlu dicatat bahwa angka ini merupakan hasil evaluasi statistik terhadap 78 sampel MIDI secara keseluruhan. Pada praktik generasi secara langsung (*real-time inference*), sistem menunjukkan performa harmonik yang jauh lebih baik berkat penerapan mekanisme *rule-based constraints* melalui fungsi *snap_to_scale* yang secara aktif memfilter token nada di luar tangga nada yang dipilih. Hal ini membuktikan bahwa arsitektur RWKV yang telah melalui 5 epoch pelatihan dengan nilai Cross-Entropy Loss sebesar 0,8752 telah mencapai konvergensi yang memadai, dan bahwa kombinasi antara model generatif dengan pembatasan berbasis aturan musik formal merupakan pendekatan yang efektif untuk menekan anomali nada pada lingkungan komputasi dengan sumber daya terbatas.

Pencapaian harmonisasi yang baik dalam penelitian ini merupakan hasil sinergi antara kemampuan generatif arsitektur RWKV yang telah terkonvergensi dalam 5 epoch pelatihan dengan mekanisme *rule-based constraints* berbasis teori musik formal. Berbeda dengan pendekatan *end-to-end* seperti Museformer[3] dan MMT [10] yang mengandalkan skala data dan epoch pelatihan yang sangat besar untuk mencapai koherensi harmoni, penelitian ini membuktikan bahwa integrasi antara model generatif ringan dengan pembatasan tangga nada berbasis aturan dapat menghasilkan melodi yang harmonis bahkan pada lingkungan komputasi terbatas. Pendekatan ini sejalan dengan temuan [12] yang menekankan pentingnya pemilihan representasi data yang tepat, dan memperluas wawasan tersebut dengan menunjukkan bahwa kontrol berbasis aturan musik formal merupakan komplemen yang efektif bagi model generatif berskala kecil.



Gambar 5. Grafik Evaluasi Nada Fales & Anomali Melodi

 https://doi.org/10.56211/sudo.v5i2.1777

KESIMPULAN DAN SARAN

Penelitian ini berhasil mengembangkan sistem generator melodi MIDI berbasis web yang mengintegrasikan kontrol modal tangga nada bagi pengguna dengan mengimplementasikan arsitektur RWKV. Hasil pengujian menunjukkan bahwa model mampu mencapai *Cross-Entropy Loss* sebesar 0,8752 dan *Perplexity* sebesar 2,40 hanya dalam 5 epoch pelatihan, jauh lebih baik dibandingkan model LSTM konvensional yang memerlukan 100 epoch untuk mencapai *perplexity* sekitar 18 [8]. Arsitektur RWKV terbukti sangat efisien dalam penggunaan sumber daya dengan penggunaan memori GPU hanya sebesar 2,1 GB, sejalan dengan keunggulan kompleksitas linier $O(T)$ yang secara teoritis jauh lebih rendah dibandingkan kompleksitas kuadratik $O(N^2)$ pada arsitektur berbasis Transformer konvensional [4,5], yang memungkinkan proses generasi melodi secara *real-time*. Temuan ini mengkonfirmasi bahwa arsitektur RWKV [6] merupakan pilihan yang tepat untuk generasi musik simbolik pada lingkungan dengan sumber daya komputasi terbatas. Prinsip Linear Attention yang diperkenalkan oleh [4] dan diimplementasikan dalam RWKV terbukti efektif dalam mempertahankan koherensi jangka panjang pada sekuens MIDI tanpa biaya komputasi kuadratik dari Transformer tradisional. Representasi MIDI dengan tokenisasi numerik yang digunakan, sebagaimana divalidasi oleh [1] dan [9], terbukti sebagai pilihan yang tepat untuk menangkap pola temporal musik secara efisien. Meskipun parameter Temperature yang lebih tinggi memberikan fleksibilitas eksplorasi yang lebih besar, anomali lompatan melodi tetap menjadi tantangan yang konsisten pada seluruh variasi Temperature. Berdasarkan keterbatasan *hardware* yang digunakan, penelitian selanjutnya disarankan untuk melakukan pelatihan dengan jumlah epoch yang lebih tinggi, menggunakan dataset yang lebih spesifik berdasarkan genre musik, serta mengeksplorasi pendekatan representasi data alternatif seperti yang diusulkan oleh [12] untuk menekan rasio nada fales dan meningkatkan kualitas musik.

DAFTAR PUSTAKA

[1] S. Ji, X. Yang, and J. Luo, "*A Survey on Deep Learning for Symbolic Music Generation: Representations, Algorithms, Evaluations, and Challenges*," *ACM Comput. Surv.*, vol. 56, no. 1, Dec. 2023, doi: 10.1145/3597493.

[2] P. S. Yadav, S. Khan, Y. V. Singh, P. Garg, and R. S. Singh, "*A Lightweight Deep Learning-Based Approach for Jazz Music Generation in MIDI Format*," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/2140895.

[3] B. Yu et al., "Museformer: Transformer with Fine-and Coarse-Grained Attention for Music Generation," in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. doi: 10.48550/arXiv.2210.10349.

[4] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "*Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention*," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 5156–5165. doi: 10.48550/arXiv.2006.16236.

[5] B. Peng et al., "*RWKV: Reinventing RNNs for the Transformer Era*," Dec. 2023, Accessed: May 2, 2026. [Online]. Available: http://arxiv.org/abs/2305.13048.

[6] B. Peng et al., "*Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence*," Sep. 2024, Accessed: May 20, 2026. [Online]. Available: http://arxiv.org/abs/2404.05892.

[7] A. Mao, M. Mohri, and Y. Zhong, "*Cross-Entropy Loss Functions: Theoretical Analysis and Applications*," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 23803–23828. doi: 10.48550/arXiv.2304.07288.

[8] T. Nagoshi, "*MORTM: MoE-Optimized Rhythmic Transformer Model for Symbolic MIDI Generation*," in *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2025. doi: 10.1109/APSIPAASC65261.2025.11248983.

[9] P. Lisena, A. Meroño-Peñuela, and R. Troncy, "*MIDI2vec: Learning MIDI embeddings for reliable prediction of symbolic music metadata*," *Semant. Web*, vol. 13, no. 3, pp. 357–377, 2022, doi: 10.3233/SW-210446.

[10] H.-W. Dong, K. Chen, S. Dubnov, J. McAuley, and T. Berg-Kirkpatrick, "*Multitrack Music Transformer*," May 2023, Accessed: May 20, 2026. [Online]. Available: http://arxiv.org/abs/2207.06983.

[11] Y. Duan et al., "*VISION-RWKV: EFFICIENT AND SCALABLE VISUAL PERCEPTION WITH RWKV-LIKE ARCHITECTURES*," 2025. doi: 10.48550/arXiv.2403.02308.

[12] X. Qu et al., "*MUPT: A Generative Symbolic Music Pre-Trained Transformer*," arXiv preprint arXiv:2404.06393, 2024. doi: 10.48550/arXiv.2404.06393.

[13] Muhammad Djamaluddin, I. Fathurrahman, and M. Nurul Wathani, "*Pengembangan Model Deep Learning Long Short-Term Memory untuk Generasi Musik Berbasis Data MIDI*," *Infotek: Jurnal Informatika dan Teknologi*, vol. 9, no. 1, pp. 197–208, Jan. 2026, doi: 10.29408/jit.v9i1.33165.

NOMENKLATUR

Simbol-simbol yang digunakan dalam rumus Weighted Kernel Value (wkv_t) adalah sebagai berikut:

wkv_t	Weighted Kernel Value pada waktu t
t	indeks waktu saat ini
I	indeks waktu masa lalu (variabel iterasi penjumlahan, i = 1, 2, ..., t−1)
w	parameter peluruhan eksponensial (decay weight), w > 0
k_i	nilai kunci (key) pada time-step ke-i
k_t	nilai kunci (key) pada time-step saat ini t
v_i	nilai (value) pada time-step ke-i
v_t	nilai (value) pada time-step saat ini t
u	parameter skalar bias pada eksponen time-step t
e	bilangan Euler (basis eksponen alami, ≈ 2,718)
Σ	notasi penjumlahan dari i = 1 hingga t−1

 https://doi.org/10.56211/sudo.v5i2.1777