INFERENCE-LEVEL OPTIMIZATION PADA SISTEM *VOICE CONVERSATIONAL* AI BERBASIS CPU MELALUI METODE *QUANTIZATION* DAN *RUNTIME MODEL CONVERSION*

SKRIPSI

Kenny Budiarso Lawson

00000081065

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS TEKNIK DAN INFORMATIKA

UNIVERSITAS MULTIMEDIA NUSANTARA

TANGERANG

2025

INFERENCE-LEVEL OPTIMIZATION PADA *SISTEM VOICE CONVERSATIONAL* AI BERBASIS CPU MELALUI METODE *QUANTIZATION* DAN *RUNTIME MODEL CONVERSION*

SKRIPSI

Diajukan sebagai Salah Satu Syarat untuk Memperoleh

Gelar Sarjana Komputer

Kenny Budiarso Lawson

00000081065

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS TEKNIK DAN INFORMATIKA

UNIVERSITAS MULTIMEDIA NUSANTARA

TANGERANG

2025

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Kenny Budiarso Lawson

Nomor Induk Mahasiswa : 00000081065

Program Studi : Sistem Informasi

Skripsi dengan judul: *INFERENCE-LEVEL OPTIMIZATION* PADA SISTEM *VOICE CONVERSATIONAL AI* BERBASIS CPU MELALUI METODE *QUANTIZATION* DAN *RUNTIME MODEL CONVERSION*

Merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 18 Desember 2025

Kenny Budiarso Lawson

LEMBAR PERNYATAAN PENGGUNAAN BANTUAN KECERDASAN ARTIFISIAL (AI)

Saya yang bertanda tangan di bawah ini:

Nama Lengkap : Kenny Budiarso Lawson

NIM : 00000081065

Program Studi : Sistem Informasi

Judul Skripsi : *Inference*-Level Optimization pada Sistem *Voice conversational* AI Berbasis CPU Melalui Metode *Quantization* dan *Runtime Model Conversion*

Dengan ini saya menyatakan secara jujur menggunakan bantuan Kecerdasan Artifisial (AI) dalam pengerjaan Tugas/Project/~~Tugas Akhir~~ sebagai berikut:

- ✓ Menggunakan AI sebagaimana diizinkan untuk membantu dalam menghasilkan ide-ide utama saja
- ✓ Menggunakan AI sebagaimana diizinkan untuk membantu menghasilkan teks pertama saja
- ✓ Menggunakan AI untuk menyempurnakan sintaksis dan tata bahasa untuk pengumpulan tugas
- ☐ Karena tidak diizinkan: Tidak menggunakan bantuan AI dengan cara apa pun dalam pembuatan tugas

Saya juga menyatakan bahwa:

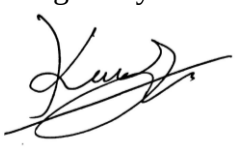
(1) Menyerahkan secara lengkap dan jujur penggunaan perangkat AI yang diperlukan dalam tugas melalui Formulir Penggunaan Perangkat Kecerdasan Artifisial (AI)

(2) Saya mengakui bahwa saya telah menggunakan bantuan AI dalam tugas saya baik dalam bentuk kata, paraphrase, penyertaan ide atau fakta penting yang disarankan oleh AI dan saya telah menyantumkan dalam sitasi serta referensi

(3) Terlepas dari pernyataan di atas, tugas ini sepenuhnya merupakan karya saya sendiri

Tangerang, 18 Desember 2025

Yang Menyatakan,



Kenny Budiarso Lawson

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan di bawah ini:

Nama Lengkap	: Kenny Budiarso Lawson
Nomor Induk Mahasiswa	: 00000081065
Program Studi	: Sistem Informasi
Jenjang	: S1
Jenis Karya	: Skripsi
Judul Karya Ilmiah	: *Inference*-Level Optimization pada Sistem *Voice conversational* AI Berbasis CPU Melalui Metode *Quantization* dan *Runtime Model Conversion*

Menyatakan dengan sesungguhnya bahwa saya bersedia* **(pilih salah satu):**

☑ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.

☐ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.

☐ Lainnya, pilih salah satu:

☐ Hanya dapat diakses secara internal Universitas Multimedia Nusantara

☐ Embargo publikasi karya ilmiah dalam kurun waktu 3 tahun.

Tangerang, 18 Desember 2025

Kenny Budiarso Lawson

* Pilih salah satu opsi tanpa harus menghapus opsi yang tidak dipilih

** Jika tidak bisa membuktikan LoA jurnal/ HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk diunggah ke KC UMN dan menjadi hak institusi UMN

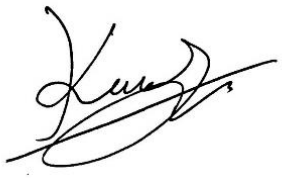
KATA PENGANTAR

Puji Syukur kepada Tuhan Yang Maha Esa atas berkat dan karunia-Nya dalam selesainya penulisan Laporan Skripsi ini dengan judul: "*Inference-Level Optimization* pada Sistem *Voice conversational* AI Berbasis CPU melalui Metode *Quantization* dan *Runtime Model Conversion*." Saya menyadari bahwa sulit untuk menyelesaikan Laporan Skripsi ini tanpa bantuan dan bimbingan dari berbagai pihak. Maka dari itu, saya mengucapkan terima kasih kepada:

1. Dr. Ir. Andrey Andoko, M.Sc., Ph.D, selaku Rektor Universitas Multimedia Nusantara.
2. Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Teknik dan Informatika Universitas Multimedia Nusantara.
3. Ibu Ririn Ikana Desanti S. Kom., M. Kom., selaku Ketua Program Studi Sistem Informasi Universitas Multimedia Nusantara.
4. Dr. Erick Fernando S. Kom., M. S. I., sebagai Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga dan teman-teman saya yang telah memberikan bantuan dukungan material dan moral dalam proses pengerjaan, sehingga penulis dapat menun*task*an tugas akhir ini.

Semoga laporan ini dapat memberikan dampak positif bagi berbagai pihak yang terlibat dalam penyusunan laporan, dan dapat memberikan kontribusi sebagai wawasan bagi pembaca, pihak kampus, serta penelitian di masa depan.

Tangerang, 18 Desember 2025



Kenny Budiarso Lawson

INFERENCE-LEVEL OPTIMIZATION PADA SISTEM *VOICE CONVERSATIONAL* AI BERBASIS CPU MELALUI METODE *QUANTIZATION* DAN *RUNTIME MODEL CONVERSION*

Kenny Budiarso Lawson

ABSTRAK

Voice conversational AI semakin dibutuhkan seiring meningkatnya kebutuhan pengguna akan interaksi yang alami, efisien, dan bebas kendali manual dalam berbagai aplikasi seperti asisten virtual, layanan pelanggan, dan sistem edukasi. Namun, arsitektur *cascaded* yang menggabungkan *Automatic Speech Recognition* (ASR), *Small Language Model* (SLM), dan *Text-to-Speech* (TTS) masih menghadapi kendala berupa *latency* tinggi dan beban komputasi besar, terutama ketika dijalankan pada lingkungan CPU-*only* yang umum digunakan pada *server* berbiaya rendah maupun perangkat lokal. Penelitian ini menerapkan teknik optimasi pada tahap *inference* tanpa melakukan pelatihan ulang model, meliputi konversi model Whisper Small menggunakan CTranslate2 dengan INT8 *quantization*, GGUF *quantization* untuk Gemma 3 1B, serta konversi VITS MMS ke ONNX dengan optimasi graf dan *static* INT8 *quantization*. Evaluasi dilakukan menggunakan *dataset* Common Voice Indonesian untuk ASR dan kumpulan *prompt* teks yang disusun secara manual untuk SLM dan TTS, dengan metrik pengujian berupa *Word Error Rate, Character Error Rate, latency, Real-time Factor, perplexity, tokens-per-second*, dan ukuran model. Hasil penelitian menunjukkan bahwa optimasi Whisper menggunakan CTranslate2 INT8 menurunkan *latency* sebesar 19,87% dengan peningkatan *error rate* yang minimal; optimasi Gemma menggunakan GGUF Q8_0 menghasilkan penurunan *latency* sebesar 44,85% dan peningkatan *throughput* 26,78%; sedangkan optimasi VITS memberikan percepatan 1,61x dengan penurunan *latency* sebesar 37,74%. Secara keseluruhan, sistem *end-to-end* mengalami penurunan *latency* sebesar 30,09% dari 19,76 detik menjadi 13,81 detik, disertai dengan reduksi ukuran model pada semua komponen, menunjukkan bahwa kombinasi teknik optimasi *inference-level* efektif meningkatkan kinerja sistem *voice conversational* AI pada lingkungan CPU-*only*.

Kata kunci: *Cascaded architecture, inference optimization, model runtime conversion, quantization, voice conversational* AI

INFERENCE-LEVEL OPTIMIZATION IN CPU-BASED CONVERSATIONAL AI VOICE SYSTEM USING QUANTIZATION AND RUNTIME MODEL CONVERSION METHOD

Kenny Budiarso Lawson

ABSTRACT (English)

Conversational AI is increasingly needed as users need natural, efficient, and manual-free interactions in various applications such as virtual assistants, customer service, and education systems. However, cascaded architectures that combine Automatic Speech Recognition (ASR), Small Language Model (SLM), and Text-to-Speech (TTS) still face obstacles in the form of high latency and large computational burdens, especially when running in CPU-only environments commonly used on low-cost servers or local devices. This study applies optimization techniques at the inference stage without retraining the model, including the conversion of the Whisper Small model using CTranslate2 with INT8 quantization, GGUF quantization for Gemma 3 1B, and the conversion of VITS MMS to ONNX with graph optimization and static INT8 quantization. The evaluation was conducted using the Indonesian Common Voice dataset for ASR and a manually compiled collection of prompt texts for SLM and TTS, with test measurements in the form of Word Error Rate, Character Error Rate, latency, Real-time Factor, perplexity, tokens-per-second, and model size. The results show that Whisper optimization using CTranslate2 INT8 reduces latency by 19.87% with minimal increase in error rate; Gemma optimization using GGUF Q8_0 results in a 44.85% decrease in latency and a 26.78% increase in throughput; while VITS optimization provides a 1.61× acceleration with a 37.74% decrease in latency. Overall, the end-to-end system experiences a 30.09% decrease in latency from 19.76 seconds to 13.81 seconds, accompanied by a reduction in model size across all components, indicating that the combination of inference-level optimization techniques effectively improves the performance of AI conversational voice systems in CPU-only environments.

***Keywords**: Cascaded architecture, inference optimization, model runtime conversion, quantization, voice conversational AI*

DAFTAR ISI

DAFTAR TABEL

DAFTAR RUMUS

DAFTAR GAMBAR

DAFTAR LAMPIRAN