202210370311251
Muhammad Hauzan Afif
Teknik Informatika

IMPLEMENTASI ASISTEN VIRTUAL AI MENGGUNAKAN FINE-TUNING MODEL LLAMA 3.2 DENGAN OLLAMA UNTUK SISTEM INFORMASI AKADEMIK PADA PLATFORM UNITY

TUGAS AKHIR

Disusun oleh:

Muhammad Hauzan Afif

(202210370311251)

Bidang Minat Game Cerdas

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MALANG

2025

202210370311251
Muhammad Hauzan Afif
Teknik Informatika

LEMBAR PERSETUJUAN

IMPLEMENTASI ASISTEN VIRTUAL AI MENGGUNAKAN FINE-TUNING MODEL LLAMA 3.2 DENGAN OLLAMAUNTUK SISTEM INFORMASI AKADEMIK PADA PLATFORM UNITY

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
Informatika Universitas Muhammadiyah Malang

Menyetujui,
Malang, *16 April 2026*

Dosen Pembimbing 1

Hardianto Wibowo S.Kom, MT.
NIP. 10816120592PNS.

LEMBAR PENGESAHAN

IMPLEMENTASI ASISTEN VIRTUAL AI MENGGUNAKAN FINE TUNING MODEL LLAMA 3.2 DENGAN OLLAMAUNTUK SISTEM INFORMASI AKADEMIK PADA PLATFORM UNITY

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
InformatikaUniversitas Muhammadiyah Malang

Disusun Oleh :

Muhammad Hauzan Afif

202210370311251

Tugas Akhir ini telah diuji dan dinyatakan lulus melalui sidang majelis penguji pada tanggal 16 April 2026

Menyetujui,

Dosen Pembimbing 1

Hardianto Wibowo S.Kom, MT.
NIP. 10816120592PNS.

Dosen Penguji 2

Lailatul Husniah S.ST., MT.
NIP. 10816120580PNS.

Dosen Penguji 1

Ali Sofyan Kholimi S.Kom., M.Kom.
NIP. 10814100562PNS.

Mengetahui,
Ketua Jurusan Informatika

Ir. Agus Eko Minarno S.Kom., M.Kom. IPM.
NIP. 10814100540PNS.

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini :

NAMA : **Muhammad Hauzan Afif**
NIM : **202210370311251**
FAK./JUR. : **Informatika**

Dengan ini saya menyatakan bahwa Tugas Akhir dengan judul **"IMPLEMENTASI ASISTEN VIRTUAL AI MENGGUNAKAN FINE-TUNING MODEL LLAMA 3.2 DENGAN OLLAMAUNTUK SISTEM INFORMASI AKADEMIK PADA PLATFORM UNITY"** beserta seluruh isinya adalah karya saya sendiri dan bukan merupakan karya tulis orang lain, baik sebagian maupun seluruhnya, kecuali dalam bentuk kutipan yang telah disebutkan sumbernya.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya ini, atau ada klaim dari pihak lain terhadap keaslian karya saya ini maka saya siap menanggung segala bentuk resiko/sanksi yang berlaku.

Mengetahui,
Dosen Pembimbing

Hardianto Wibowo S.Kom, MT.

Malang, 16 April 2026
Yang Membuat Pernyataan

Muhammad Hauzan Afif

202210370311251
Muhammad Hauzan Afif
Teknik Informatika

ABSTRAK

Sistem informasi akademik di perguruan tinggi seringkali kekurangan fitur pencarian interaktif dan *real-time*, sehingga menghambat akses mahasiswa terhadap layanan esensial dan menambah beban kerja administratif staf. Untuk mengatasi urgensi tersebut, penelitian ini bertujuan mengembangkan dan mengevaluasi asisten virtual kecerdasan buatan generatif menggunakan model Llama 3.2 yang di-*fine-tuning* dan diintegrasikan ke platform Unity melalui Ollama. Penelitian ini menggunakan metode *Research and Development* dengan menerapkan *Parameter-Efficient Fine-Tuning* (PEFT) melalui algoritma *Quantized Low-Rank Adaptation* (QLoRA) 4-bit untuk mengadaptasi model berparameter tiga miliar menggunakan 8.725 data percakapan akademik. Model dieksekusi secara lokal menggunakan Ollama yang bertindak sebagai jembatan REST API dengan antarmuka 3D di Unity. Hasil pengujian menunjukkan proses *fine-tuning* berhasil menghasilkan model yang sangat kontekstual dengan lingkungan akademik institusi, mencatatkan titik konvergensi akhir *Training Loss* di angka 1,3494. Evaluasi performa bahasa mencatat nilai *Perplexity* keseluruhan sebesar 10,67, yang mengindikasikan tingkat kepastian linguistik yang sangat baik. Secara komputasi, sistem beroperasi stabil dengan latensi rata-rata 5,43 detik per kueri dan kecepatan pemrosesan 20,42 token per detik pada perangkat keras lokal skala menengah, melampaui rata-rata kecepatan kognitif membaca manusia. Lebih lanjut, rata-rata *ROUGE-L* sebesar 15,34% dikombinasikan dengan perolehan skor kepuasan *User Acceptance Testing* (UAT) sebesar 4,75 dari skala 5,00 memvalidasi keandalan sistem dalam melakukan parafrase bahasa yang sangat natural layaknya manusia tanpa menimbulkan halusinasi fakta. Meskipun sistem ini sangat efisien secara lokal, kelemahan saat ini terletak pada ketergantungan mutlak pada *prompt* berbahasa Indonesia dan memori data yang masih bersifat statis. Penelitian di masa depan disarankan untuk mengeksplorasi *fine-tuning* multibahasa dan integrasi *Retrieval-Augmented Generation* (RAG) untuk menangani pembaruan data akademik yang dinamis.

Kata Kunci: Asisten Virtual, Kecerdasan Buatan Generatif, Llama 3.2, Low-Rank Adaptation, Unity Engine.

202210370311251
Muhammad Hauzan Afif
Teknik Informatika

ABSTRACT

Academic information systems in higher education often lack interactive, real-time query capabilities, hindering student access to essential services and increasing the administrative burden on staff. To address this urgency, this study aims to develop and evaluate a generative artificial intelligence virtual assistant using a fine-tuned Llama 3.2 model integrated into the Unity platform via Ollama. A Research and Development approach was employed, utilizing Parameter-Efficient Fine-Tuning (PEFT) through 4-bit Quantized Low-Rank Adaptation (QLoRA) to adapt a three-billion-parameter model using 8,725 academic conversation datasets. The model was deployed locally using Ollama, acting as a REST API bridge to a 3D interface in Unity. The results demonstrated that the fine-tuning process successfully yielded a highly contextualized model tailored for academic environments, achieving a final Training Loss convergence of 1.3494. Linguistic performance evaluation recorded an overall Perplexity score of 10.67, indicating excellent linguistic certainty. Computationally, the system operated stably with an average latency of 5.43 seconds per query and a generation speed of 20.42 tokens per second on mid-range local hardware, surpassing average human cognitive reading speeds. Furthermore, an average ROUGE-L of 15.34% combined with a User Acceptance Testing (UAT) satisfaction score of 4.75 out of 5.00 validated the system's reliability in performing highly natural, human-like language paraphrasing without factual hallucinations. While the system is highly efficient locally, current limitations include a strict reliance on Indonesian language prompts and static data memory. Future research should explore multilingual fine-tuning and the integration of Retrieval-Augmented Generation (RAG) to handle dynamic academic data updates.

Keywords: Generative AI, Llama 3.2, Low-Rank Adaptation, Unity Engine, Virtual Assistant

DAFTAR ISI

LEMBAR ii.
PERSETUJUAN iii
LEMBAR PENGESAHAN iv
LEMBAR PERNYATAAN v
ABSTRAK vi
ABSTRACT vii
DAFTAR ISI 8
BAB 1 8
PENDAHULUAN 8
1.1 Latar Belakang 8
1.2 Rumusan Masalah 9
1.3 Tujuan Penelitian 10
1.4 Manfaat Penelitian 10
1.5 Batasan Penelitian 11
BAB II 13
TINJAUAN PUSTAKA DAN LANDASAN TEORI 13
2.1 Large Language Models(LLMs) dan Arsitektur Transformer 13
2.2 Llama 3.2 13
2.3 Tantangan Full Fine-Tuning 14
2.4 Parameter-Efficient Fine-Tuning(PEFT) 15
2.5 Low-Rank Adaptation(LoRA) 16
2.6 Ollama dan Manajemen Model Lokal 17
2.7 Unity Engine dan integrasi API 17
2.8 Penelitian Terdahulu dan Posisi Penelitian 18
BAB III 19
METODOLOGI PENELITIAN 19
3.1 Pendekatan Penelitian 19
3.2 Arsitektur Sistem 20
3.3 Tahapan Penelitian 21
3.3.1 Tahap 1:Persiapan Lingkungan dan Pengolahan Data 21
3.3.2 Tahap 2:Proses *Fine-Tuning* Teknis Menggunakan LoRA 23

3.3.3 Tahap 3: Implementasi dan *Deployment* Lokal 25
3.4 Metode Evaluasi Kinerja Model 26
3.5 Tabel Pengerjaan 28
BAB IV 31
HASIL DAN PEMBAHASAN 31
4.1 Hasil Skenario Pelatihan dan Proses Fine-Tuning 31
4.2 Evaluasi Waktu Respons, Kecepatan Komputasi, dan Skor ROUGE (*Latency, Speed, & ROGUE*) 33
4.3 Evaluasi Pemahaman Bahasa: Perplexity dan Skor ROUGE 40
4.3.1 Analisis *Perplexity* (PPL) 40
4.3.2 Analisis ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) 42
4.4 Evaluasi Kualitatif Berpusat pada Manusia (*Human-Centered Evaluation*) 43
BAB V 45
KESIMPULAN DAN SARAN 45
5.1 Kesimpulan 45
5.2 Saran 46
DAFTAR PUSTAKA 47

DAFTAR PUSTAKA

1. A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.

2. T. Wolf *et al.*, "HuggingFace's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

3. Meta, "Llama 3.2 model card," Hugging Face, Sep. 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

4. V. Lialin, V. Deshpande, and A. Rumshisky, "A survey on parameter-efficient fine-tuning," *arXiv preprint arXiv:2303.15647*, 2023.

5. E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

6. J. Primordial, G. Bgeron, and The Ollama Team, "Ollama," 2024. [Online]. Available: https://github.com/ollama/ollama

7. Z. Zhang, Y. Wang, Z. Liu, and J. Zhang, "Cross-modal AIGC integration in Unity3D for enhanced game art generation," *Electronics*, vol. 14, no. 6, p. 1101, Mar. 2025.

8. L. von Werra, Y. Belkada, K. Tunstall, E. Raffel, and The Hugging Face Team, "TRL: Transformer reinforcement learning," 2020. [Online]. Available: https://github.com/huggingface/trl

9. T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023.

10. S. Raschka, Y. Belkada, and The Hugging Face Team, "PEFT: Parameter-efficient fine-tuning," 2022. Accessed: Feb. 28, 2026. [Online]. Available: https://github.com/huggingface/peft

11. Z. Zhang *et al.*, "A survey on evaluation of large language models," *ACM*

Transactions on Intelligent Systems and Technology, vol. 15, no. 3, pp. 1–45, 2024.

12. I. Mansha, "Resource-efficient fine-tuning of LLaMA-3.2-3B for medical chain-of-thought reasoning," *arXiv preprint arXiv:2510.05003*, 2025.

13. S. Minaee *et al.*, "Large language models: A survey," *arXiv preprint arXiv:2402.06196*, 2024.

14. N. Ghosh, A. Alistarh, and C. C. Can, "LoRA+: Efficient low rank adaptation of large models," *arXiv preprint arXiv:2402.12354*, 2024.

15. S. Liu, R. Zhang, J. Leskovec, and P. H. S. Torr, "DoRA: Weight-decomposed low-rank adaptation," *arXiv preprint arXiv:2402.09353*, 2024.

16. R. Taori *et al.*, "Stanford Alpaca: An instruction-following LLaMA model," *Stanford Center for Research on Foundation Models*, 2023.

17. H. Touvron *et al.*, "LLaMA: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

18. Y. Hu *et al.*, "LoRS: Efficient low-rank adaptation for sparse large language model," *arXiv preprint arXiv:2501.08582*, 2025.

19. Unsloth, "LoRA hyperparameters guide," Unsloth Documentation, 2024. [Online]. Available: https://docs.unsloth.ai/get-started/fine-tuning-llms-guide/lora-hyperparameters-guide

20. J. Zhao, M. Xu, and Z. Liu, "Efficient fine-tuning of large language models using LoRA variants for domain-specific adaptation," *IEEE Access*, vol. 12, pp. 45512–45524, 2024.

21. R. Singh, A. Patel, and K. Suresh, "Lightweight parameter-efficient fine-tuning for on-device LLM deployment in low-resource environments," in *Proc. IEEE Int. Conf. Artificial Intelligence and Computer Applications (ICAICA)*, 2024, pp. 210–216.

22. H. Chen, L. Yu, and F. Sun, "Instruction-tuned LLaMA models for task-oriented virtual assistants: A comparative study," *ACM Trans. Asian Low-Resource Language Information Processing (TALLIP)*, vol. 24, no. 3, pp. 1–19, 2025.

23. D. Park, Y. Han, and T. Choi, "Local deployment of LLM-based chatbots using quantization and PEFT on edge GPUs," in *Proc. IEEE Int. Conf. Cloud Computing and Edge Intelligence (CCEI)*, 2024, pp. 98–104.

24. A. Rahman and S. Nugroho, "Pengembangan asisten virtual akademik berbasis large language model dengan integrasi platform interaktif," *Jurnal Teknologi Informasi*

dan Komunikasi (JTIK), vol. 9, no. 2, pp. 134–145, 2024.

25. G. Patel, Y. Wang, and T. Nguyen, “Unity-based interactive AI assistants using RESTful LLM backends for educational systems,” in *Proc. ACM Symp. Virtual Reality Software and Technology (VRST)*, 2024, pp. 300–308.

26. P. Lewis *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

27. J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

28. J. Lee and M. Kim, “Building an LLM development environment in Unity for developing conversational virtual avatars,” *Proceedings of the HCI Society of Korea*, pp. 120–125, 2024.

UNIVERSITAS
MUHAMMADIYAH
MALANG

FAKULTAS TEKNIK

INFORMATIKA
informatika.umm.ac.id | informatika@umm.ac.id

FORM CEK PLAGIARISME LAPORAN TUGAS AKHIR

Nama Mahasiswa : **Muhammad Hauzan Afif**

NIM : **202210370311251**

Judul TA : IMPLEMENTASI ASISTEN VIRTUAL AI MENGGUNAKAN FINE-TUNING MODEL LLAMA 3.2 DENGAN OLLAMAUNTUK SISTEM INFORMASI AKADEMIK PADA PLATFORM UNITY

Hasil Cek Plagiarisme dengan Turnitin

No.	Komponen Pengecekan	Nilai Maksimal Plagiarisme (%)	Hasil Cek Plagiarisme (%) *
1.	Bab 1 – Pendahuluan	10 %	4%
2.	Bab 2 – Daftar Pustaka	25 %	0%
3.	Bab 3 – Analisis dan Perancangan	25 %	0%
4.	Bab 4 – Implementasi dan Pengujian	15 %	0%
5.	Bab 5 – Kesimpulan dan Saran	5 %	5%
6.	Makalah Tugas Akhir	20%	4%

) Hasil cek plagiarism diisi oleh pemeriksa (staf TU)

***) Maksimal 5 kali (4 Kali sebelum ujian, 1 kali sesudah ujian)**

Mengetahui,

Pemeriksa (Staff TU)

(.....)