Arsitektur *Chatbot* WhatsApp Hibrida *Rasa-DeepSeek*: Desain dan Evaluasi Performa

WhatsApp Hybrid Chatbot Architecture Rasa-DeepSeek: Design and Performance Evaluation

[1]Iqbaluddin Syam Had*, [2]Fandy Setyo Utomo, [3]Giat Karyono, [4]Dwi Putriana Nuramanah Kinding
[1,2,3]Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto
[4]Program Studi Agribisnis, Fakultas Pertanian, Universitas Jenderal Soedirman
[1,2,3]Jl. Letjend Pol. Soemarto No.127, Watumas, Purwanegara, Kec. Purwokerto Utara, Banyumas, Jawa Tengah, Indonesia
[4]Jl. DR. Soeparno No.63, Karang Bawang, Grendeng, Kec. Purwokerto Utara, Banyumas, Jawa Tengah, Indonesia
*e-mail: iqbaluddinsh@gmail.com

(***received*:** 11 October 2025, ***revised*:** 5 January 2026, ***accepted*:** 6 January 2026)

Abstrak

Penelitian ini merancang dan menguji *chatbot* hibrida untuk domain spesifik dengan dua masalah utama, yakni cakupan *NLU* yang terbatas dan variabilitas latensi serta biaya ketika seluruh pertanyaan diarahkan ke *LLM*. Solusi yang diusulkan menggabungkan jalur deterministik berbasis *Rasa* dengan *fallback DeepSeek*, dimana *Rasa* menangani *NLU*, *rules*, *stories*, serta penyimpanan konteks mk dan jk, sedangkan *LLM* hanya dipanggil saat *threshold NLU* rendah. Metode mencakup implementasi *end-to-end* melalui *bridge Node* ke *Rasa*, uji fungsional untuk memverifikasi alur *intent–entity–action*, serta uji performa berbentuk *shape* pada dua jalur akses *Rasa REST* dan *Node* ke *Rasa*, sementara jalur *LLM* diprofilkan terpisah melalui *action instrument*. Hasil menunjukkan percakapan domain terjawab dari pengetahuan terkurasi, kedua jalur memenuhi *SLO* dengan median latensi sekitar 32 milidetik tanpa *error*. Penelitian ini berkontribusi dengan menunjukkan bahwa arsitektur *chatbot* hibrida yang memisahkan jalur deterministik dan generatif mampu menjaga SLO pada domain spesifik, sekaligus mengungkap keterbatasan *LLM* dalam memahami ontologi domain sehingga menegaskan kebutuhan *semantic guardrail*.

Kata kunci: *chatbot, deepseek*, pemrosesan bahasa alami, personalisasi, *rasa framework*

Abstract

This study designed and evaluated a hybrid chatbot for a domain-specific application by addressing two main issues: limited NLU coverage and the variability of latency and cost when all queries are routed directly to an LLM. The proposed solution integrates a deterministic Rasa-based pipeline with a DeepSeek fallback mechanism. In this architecture, Rasa handles NLU processing, rules, stories, and context storage for mk *and* jk*, while the LLM is only invoked when the NLU confidence score falls below a defined threshold. The methodology includes end-to-end implementation through a Node.js bridge connected to Rasa, functional testing to validate the intent–entity–action flow, and performance testing using load (stress) testing across two access paths: the Rasa REST endpoint and the Node-to-Rasa bridge. Meanwhile, the LLM pipeline was profiled separately through instrumented action calls. The results indicate that domain-specific conversations were successfully answered using curated knowledge, and both deterministic access paths met the service level objective (SLO), achieving a median latency of approximately 32 milliseconds with no observed errors. This study contributes by demonstrating that a hybrid chatbot architecture separating deterministic and generative pipelines can maintain SLO compliance in domain-specific settings. In addition, it highlights limitations of LLMs in understanding domain ontologies, reinforcing the need for semantic guardrails.*

Keywords: *chatbot, deepseek, natural language processing, personalization, rasa framework*

http://sistemasi.ftik.unisi.ac.id

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Perkembangan teknologi kecerdasan buatan (AI) dan pemrosesan bahasa alami (NLP) telah membawa manfaat yang signifikan dalam pengembangan sistem *chatbot* yang dapat meniru percakapan manusia di era sekarang [1]. Dengan memanfaatkan *conversational artificial intelligence (CAI) framework* dan *large language models (LLMs)* yang mutakhir membuat *chatbot* tidak hanya dapat digunakan untuk tujuan dan topik yang umum, tetapi juga dapat diadaptasi untuk domain spesifik seperti pada bidang kesehatan, *e-commerce, customer service*, pendidikan, dan sebagainya [2][3][4]. Dalam konteks ini, integrasi antara *framework* berbasis aturan *(rule-based)* seperti *Rasa* dan *model* bahasa generatif seperti *DeepSeek* menawarkan solusi yang komprehensif dan lebih personalisasi untuk menangani berbagai jenis pertanyaan, mulai dari pertanyaan yang spesifik pada domain tertentu hingga pertanyaan yang lebih kompleks [5].

Akurasi pemahaman *chatbot* terhadap *utterance* pengguna merupakan tantangan signifikan dalam pengembangan *chatbot* [6]. *Rasa framework* sebagai alat untuk membuat manajemen percakapan berbasis *machine learning* [7] hadir menjadi solusi yang unggul dibandingkan dengan solusi serupa seperti Amazon Lex dan DialogFlow karena sifatnya yang dapat disesuaikan, *open source*, dan integrasi *deep learning* yang kuat [8]. *Rasa Conversational AI* terdiri dari dua komponen yaitu *Rasa Natural Language Understanding (NLU)* dan *Rasa Core* yang membuat *Rasa* memiliki performa yang baik [9][10].

Tantangan utama pada *NLU* domain spesifik adalah keterbatasan variasi data pelatihan, yang akan berdampak pada kemampuan *model* untuk secara akurat memahami respons pengguna [11]. Penerapan *LLMs* dapat membantu *chatbot* memiliki lebih banyak pengetahuan sehingga mampu merespons pertanyaan diluar cakupan data pelatihan *Rasa Framework* [2]. *LLM* yang digunakan adalah *DeepSeek-V3.2-Exp*, dipilih karena efisien pada penggunaan di sisi *serving*, mengadopsi *Multi-head Latent Attention* dan *DeepSeekMoE*, serta didukung kernel *FlashMLA* dan *DeepSeek Sparse Attention* yang mengoptimalkan dekode dan pengelolaan *KV-cache* pada konteks panjang [12]. Selain itu, *model* ini tersedia dengan harga API yang kompetitif sehingga sesuai untuk peran *fallback* berbasis API [13].

Signifikansi penelitian ini terletak pada perancangan dan evaluasi arsitektur *chatbot* hibrida yang memisahkan jalur deterministik dan jalur generatif dalam konteks domain spesifik. Sistem yang diusulkan menempatkan *Rasa* sebagai jalur utama untuk menangani *intent understanding*, *dialog management*, serta respons berbasis *curated knowledge*. *LLM* hanya dipanggil secara selektif ketika *confidence NLU* rendah. Pemisahan ini tidak hanya bertujuan untuk menekan biaya dan variabilitas latensi, tetapi juga memungkinkan evaluasi performa jalur deterministik dilakukan secara objektif tanpa bias karakteristik *LLM*, yang pada *multi-turn dialogue* diketahui menambah *overhead* pengelolaan *key–value cache* dan berpotensi memicu *head-of-line blocking* yang berdampak pada *tail latency* dan *throughput* [14]. Pendekatan yang diusulkan dalam penelitian ini berkontribusi untuk menunjukkan bahwa arsitektur *chatbot* hibrida dapat dirancang dengan jalur deterministik yang tetap memenuhi *service level objective (SLO)*, sekaligus mengungkap keterbatasan jalur generatif dalam memahami ontologi domain, sehingga menegaskan pentingnya *semantic guardrail* dan strategi evaluasi yang terpisah pada pengembangan *chatbot* hibrida domain-spesifik.

2 Tinjauan Literatur

Implementasi *Rasa framework* pada sistem *question-answering chatbot* telah dilakukan oleh beberapa peneliti sebelumnya di berbagai *platform* percakapan seperti Telegram [15][16], Facebook [17], dan bisa diaplikasikan pula pada WhatsApp [18]. Penelitian-penelitian tersebut mengintegrasikan *Rasa framework* sebagai kerangka CAI untuk menjawab pertanyaan pada domain khusus. Hasilnya *chatbot* yang dibuat dapat merespons pertanyaan dengan baik oleh pengguna, namun terjadi beberapa kesalahan hasil jawaban oleh chatbot [15].

Studi [2] meneliti bagaimana *LLMs* seperti GPT-4 dapat meningkatkan performa *chatbot* berbasis *pipeline* dengan mengatasi keterbatasan dalam merespons pertanyaan di luar cakupan. Dalam konteks *Rasa framework*, *LLMs* dapat digunakan untuk menghasilkan data pelatihan, menyempurnakan ekstraksi entitas dan sinonim, serta memperbaiki dialog melalui fitur seperti koreksi otomatis dan disambiguasi. Pendekatan *hybrid* ini memungkinkan *chatbot* berbasis *pipeline* untuk

 ISSN:2302-8149
e-ISSN:2540-9719

tetap mempertahankan privasi dan integrasi pada domain spesifik sambil memanfaatkan kemampuan *LLMs* untuk meningkatkan interaksi percakapan secara lebih alami dan fleksibel.

Penelitian [5] mengembangkan sistem *chatbot* berbasis *Rasa framework* yang diperkuat dengan *model LLMs*, seperti GPT-3.5-turbo, untuk meningkatkan fleksibilitas dalam memahami percakapan pengguna. Integrasi ini memungkinkan sistem untuk menangani pertanyaan pada topik yang umum secara lebih dinamis, sementara *Rasa* tetap berfungsi sebagai inti *NLU* yang mengelola *intent* dan entitas secara terstruktur. Dengan pendekatan ini, *chatbot* dapat menggabungkan respons yang lebih terpersonalisasi dan berbasis pengetahuan terkurasi, memberikan keseimbangan antara pemrosesan *intent* berbasis aturan dan kemampuan generatif dari *LLMs*.

Sejalan dengan tren terkini, QAS *Chatbot* pada penelitian ini mengadopsi arsitektur hibrida yang menggabungkan *Rasa* dan *model* generatif yang diimplementasikan pada aplikasi WhatsApp, dimana WhatsApp adalah aplikasi percakapan yang paling populer bahkan di seluruh dunia [19]. Penelitian ini mengevaluasi kinerja jalur deterministik pada dua arsitektur akses, yaitu *Rasa REST* langsung dan WhatsApp *bridge*, menggunakan *load profile* yang identik. Fokus pengukuran mencakup *latency percentiles*, *max latency*, dan matrik sumber daya pada *bridge*, sementara evaluasi *fallback LLM* dijalankan terpisah agar tidak bias terhadap hasil performa jalur deterministik.

3 Metode Penelitian

Chatbot kontekstual penelitian ini menggunakan domain spesifik STIFIn sebagai sumber pengetahuan. STIFIn adalah sebuah konsep yang digunakan untuk menentukan mesin kecerdasan dan lapisan otak dominan yang dimiliki seseorang [20]. Tes STIFIn menawarkan pemahaman lebih dalam tentang potensi diri, membantu pengembangan karir, serta meningkatkan komunikasi dan hubungan interpersonal. Dengan mengintegrasikan *Rasa framework* dan *DeepSeek model*, *chatbot* dirancang untuk memberikan respons yang akurat dan personalisasi berdasarkan konsep STIFIn, sekaligus memanfaatkan *LLMs* untuk meningkatkan interaksi pengguna. Pendekatan ini sejalan dengan tren terkini dalam pengembangan *chatbot* yang menekankan personalisasi dan adaptabilitas [21].

Penelitian ini mengadopsi pendekatan pengembangan sistem *(system development)* untuk merancang dan mengimplementasikan WhatsApp *chatbot* kontekstual yang mengintegrasikan *Rasa Framework* dan *DeepSeek model*. Metode penelitian terdiri dari beberapa tahap utama, yaitu: (1) analisis kebutuhan, (2) perancangan arsitektur sistem, (3) implementasi dan integrasi, (4) pengujian dan evaluasi. Setiap tahap dijelaskan secara rinci sebagai berikut.

3.1 Analisis Kebutuhan

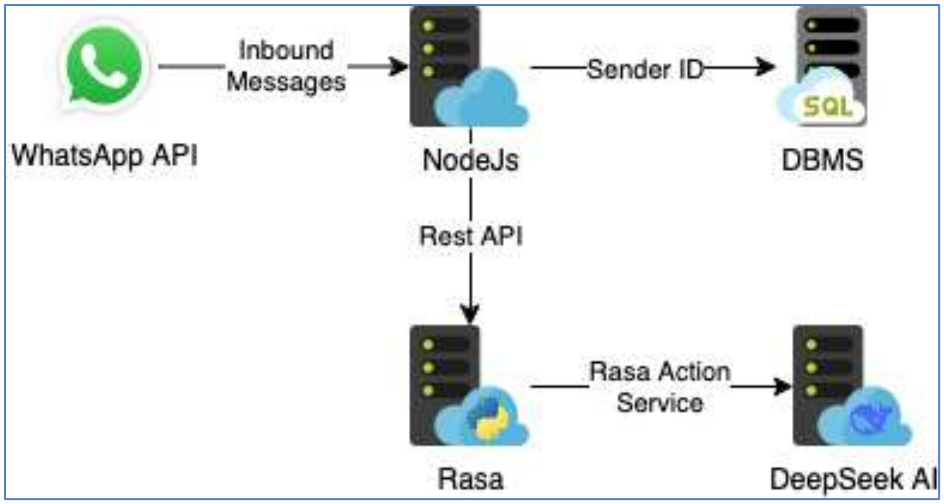
Tahap ini bertujuan untuk mengidentifikasi kebutuhan fungsional dan *non*-fungsional dari *chatbot*. Fokus utama adalah pada domain STIFIn, yang mencakup cara komunikasi, minat, bakat, dan gaya belajar berdasarkan teori otak dominan. Analisis kebutuhan dilakukan melalui pengumpulan data bersama Dwi Putriana Nuramanah Kinding, S.P., M.Si., sebagai promotor STIFIn berlisensi untuk memastikan akurasi konten. Selain itu, kebutuhan teknis seperti integrasi dengan WhatsApp API, penggunaan *Rasa Framework* untuk manajemen dialog, dan pemanfaatan *DeepSeek model* untuk pemrosesan *query* kompleks juga dianalisis.

3.2 Perancangan Arsitektur Sistem

Arsitektur sistem dirancang untuk menggabungkan kekuatan *Rasa Framework* dan *DeepSeek*. *Rasa Framework* digunakan untuk menangani alur dialog spesifik, sementara *DeepSeek* bertugas untuk menangani *query* yang memerlukan pemahaman bahasa alami mendalam yang tidak dapat dipahami oleh *Rasa framework*. Arsitektur sistem terdiri dari tiga komponen utama:

1. *Frontend*: Whatsapp sebagai antarmuka pengguna.
2. *Backend*: Aplikasi NodeJs sebagai *message listener* dan konfigurasi basis data, *Rasa Framework* untuk manajemen dialog kontekstual.
3. NLP *Model*: *DeepSeek* untuk pemrosesan *query* kompleks dan merespons pertanyaan diluar pengetahuan *Rasa*.

 ISSN:2302-8149
e-ISSN:2540-9719

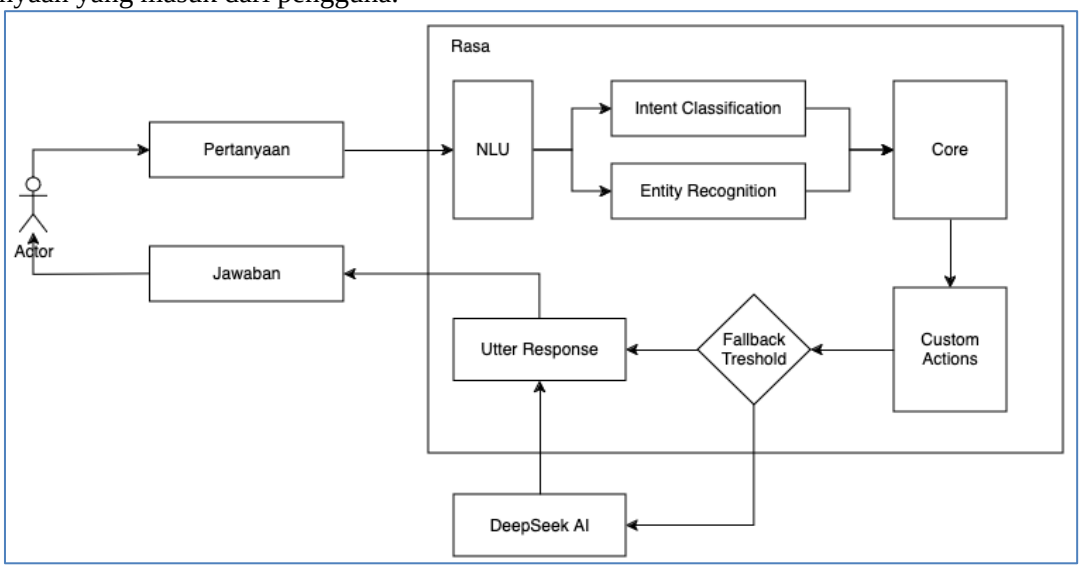


Gambar 1 Arsitektur sistem *chatbot*

3.3 Implementasi dan Integrasi

Pada tahap ini, komponen-komponen sistem diimplementasikan dan diintegrasikan. *Rasa Framework* dikonfigurasi untuk membuat klasifikasi *intent*, *entity recognition*, dan *action* yang relevan dengan domain STIFIn. Data pelatihan untuk *Rasa* disiapkan berdasarkan skenario percakapan yang telah dianalisis. *DeepSeek* diintegrasikan sebagai layanan eksternal yang dipanggil ketika *query* pengguna memerlukan pemrosesan bahasa alami yang lebih kompleks.

Gambar 2 menunjukkan alur kerja bagaimana kombinasi *Rasa* dan *model* generatif memproses pertanyaan yang masuk dari pengguna.



Gambar 2 Alur Kerja sistem *chatbot*

3.4 Pengujian dan Evaluasi

Pengujian dan evaluasi difokuskan pada pengujian fungsional dan kinerja *end-to-end chatbot* WhatsApp untuk pertanyaan berbasis STIFIn yang mengintegrasikan *Rasa* untuk *intent* dan aksi terstruktur, serta *DeepSeek* untuk *query* kompleks melalui *custom action*. *Load test* dijalankan 3 menit per-*channel*, yaitu *Rasa REST* dan WhatsApp *bridge* dengan pola 60 detik dan 120 detik. Matrik yang dicatat meliputi latensi *end-to-end*, *error/timeout rate*, serta konsumsi sumber daya per komponen *Rasa Server* dan WhatsApp *bridge*. `Kriteria kelulusan (SLO) ditetapkan p95 ≤ 120 ms dan` *error/timeout* < 1%, dengan sistem stabil pada sumber daya yang tersedia.

 ISSN:2302-8149
e-ISSN:2540-9719

4 Hasil dan Pembahasan

Chatbot kontekstual dirancang dengan berbagai skenario percakapan pada domain STIFIn. Tabel 1 menunjukkan sampel dari total 28 jenis *intent* yang digunakan pada *chatbot*. Slot digunakan untuk mengingat *entity* dari percakapan sebelumnya agar *chatbot* mampu memahami konteks yang sedang dibahas pengguna.

Tabel 1 Sampel daftar pelatihan *intent*

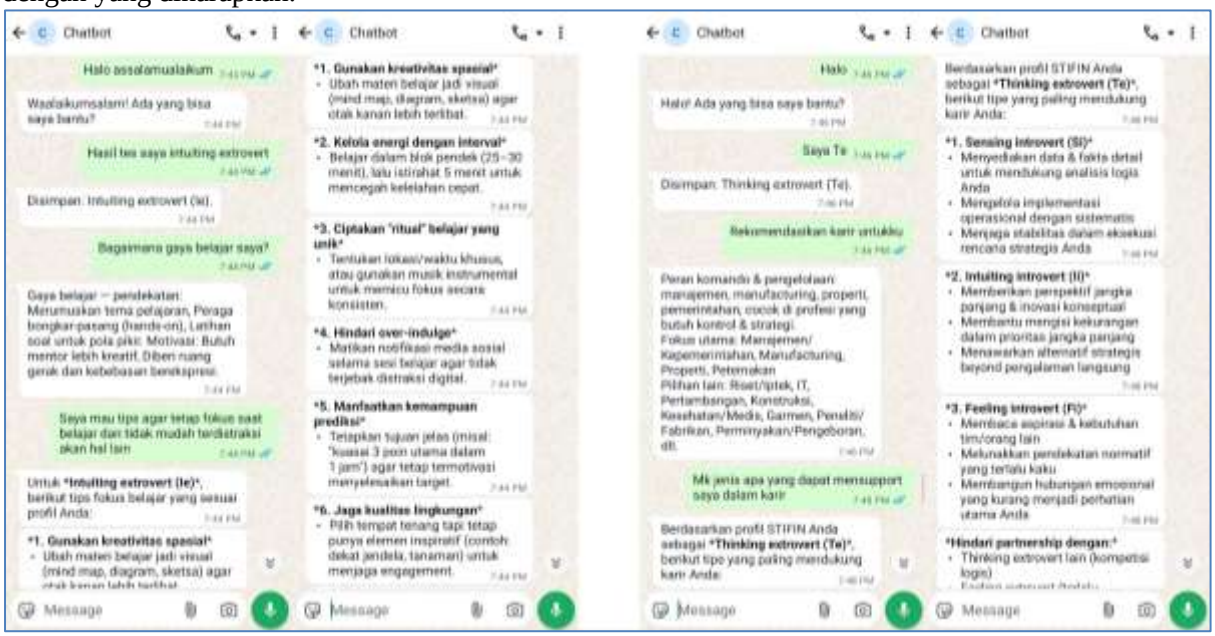
Intent	Deskripsi	Jumlah Pertanyaan
greetings	Query sapaan seperti “Halo”	5
salam_greetings	Sapaan dalam ungkapan salam seperti “Assalamu’alaikum”	2
ask_booking	Permintaan pemesanan tes STIFIn pada lokasi tertentu	4
ask_personality	Pertanyaan kepribadian dari mesin kecerdasan (MK) tertentu	3
ask_career	Pertanyaan rekomendasi karir dari mesin kecerdasan (MK) tertentu	6
ask_learning_style	Pertanyaan terkait gaya belajar dari mesin kecerdasan (MK) tertentu	3

4.1 Implementasi Chatbot

Implementasi *chatbot* hibrida berjalan *end-to-end* sesuai rancangan pada Gambar 1. *Channel* WhatsApp melalui Node.js *bridge* meneruskan pesan ke *Rasa REST*, sementara *Rasa* mengelola percakapan melalui *pipeline NLU*, *policies* (*stories/rules*), dan *action server* yang mengakses pengetahuan STIFIn dengan penyimpanan konteks mk dan jk di *tracker*. Jalur deterministik aktif untuk pertanyaan berbasis pengetahuan yang telah dilatih, dan *fallback* generatif diaktifkan selektif saat kepercayaan *NLU* rendah melalui rangkaian *utter_fallback* dan *action_llm_fallback*. Konfigurasi kunci yang digunakan pada pengujian meliputi ambang *fallback NLU*, pemanggilan *LLM* tanpa *streaming*, serta persistensi slot mk dan jk agar pengguna tidak perlu menyebut tipe berulang. *Environment* pengujian menggunakan *Rasa 3.6.21*, *Node v22.12.0*, dan *DeepSeek 3.2-Exp* dengan *keep-alive* di sisi *bridge*.

4.2 Hasil Pengujian Fungsional

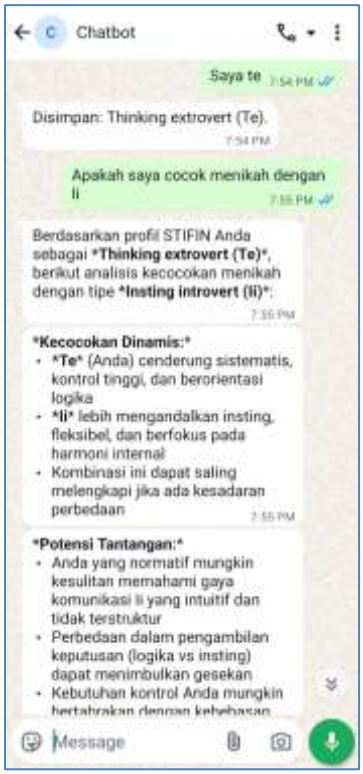
Pengujian fungsional dilakukan untuk memastikan bahwa *intent*, *entity*, dan *action* bekerja sesuai dengan yang diharapkan.



Gambar 3 Hasil pengujian fungsional

 ISSN:2302-8149
e-ISSN:2540-9719

Gambar 3 menunjukkan kemampuan *chatbot* menjawab sesuai pengetahuan yang telah dilatih *Rasa* dan mengembalikan *fallback* ke *LLM* untuk pertanyaan yang lebih kompleks dan di luar pengetahuan *Rasa*. Hal ini menunjukkan bahwa *pipeline* hibrida berjalan sesuai rancangan. *NLU* mengklasifikasikan *intent* secara tepat, mengekstraksi entitas serta mengisi slot mk dan jk secara konsisten, kemudian *policy* dan *stories* memicu *action* yang relevan sehingga jawaban deterministik dari basis pengetahuan STIFIN tersaji dengan konteks yang benar. Ketika tingkat kepercayaan *NLU* menurun, mekanisme *fallback* diaktifkan secara selektif dan tetap membawa konteks percakapan yang telah tersimpan, sehingga respons generatif tetap relevan terhadap profil pengguna. Secara fungsional, hasil ini memvalidasi alur *intent–entity–action*, ketahanan penyimpanan konteks antar giliran, serta integrasi *guardrail* yang mencegah ketergantungan berlebihan pada *LLM* sekaligus menjaga koherensi dan akurasi jawaban pada domain STIFIn.



Gambar 4 Kegagalan *LLM* menerjemahkan singkatan

Gambar 4 menyoroti batasan jalur generatif ketika ontologi domain tidak dipatuhi sepenuhnya. *LLM* keliru menafsirkan kode Ii sebagai Insting introvert, sedangkan dalam terminologi STIFIn Ii adalah Intuiting introvert, sedangkan Insting dikodekan In dan tidak memiliki jenis kemudi introvert atau extrovert karena bersumber dari dominasi otak tengah. Temuan tersebut menegaskan pentingnya pengendali semantik pada jalur *LLM*, antara lain normalisasi entitas, aturan validasi slot yang menolak kombinasi tidak sah seperti pemberian jk pada tipe Insting, serta kebijakan disambiguasi yang meminta klarifikasi ketika keluaran melanggar ontologi. Pada sisi *prompting*, penegasan ruang lingkup tertutup terhadap lima tipe STIFIN beserta kode dan padanan sah, ditambah contoh beranotasi untuk kasus Ii dan In, membantu mengurangi salah tafsir. Dengan kombinasi *guardrail* tersebut, jalur generatif tetap bermanfaat tanpa mengorbankan konsistensi taksonomi domain.

4.3 Hasil Pengujian Performa

Pengujian dilakukan menggunakan Artillery dengan *load profile* 60 detik @5rps, dilanjutkan 120 detik @20rps. Dua jalur dibandingkan antara:

1. *Rasa REST* langsung ke *endpoint* /webhooks/rest/webhook.
2. *Node* menuju *Rasa* melalui *bridge* Node.js tanpa mengirim pesan WhatsApp. *Endpoint Node* meneruskan *request* ke *Rasa* dengan HTTP *keep-alive* dan *pooling*.

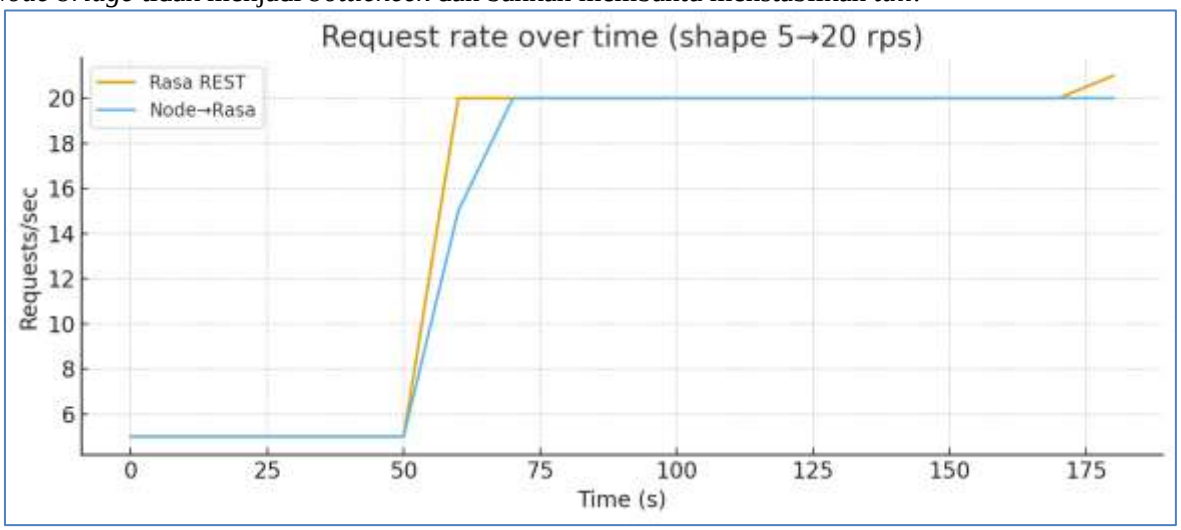
Fallback LLM dinonaktifkan agar metrik merepresentasikan jalur deterministik (*intent* menuju *rules/stories* dan *actions*). Pengujian dijalankan di mesin pengembangan lokal dengan sistem operasi Mac OS, *Node* v22.12.0, CPU Apple Silicon M1 dan RAM 8GB.

 ISSN:2302-8149
e-ISSN:2540-9719

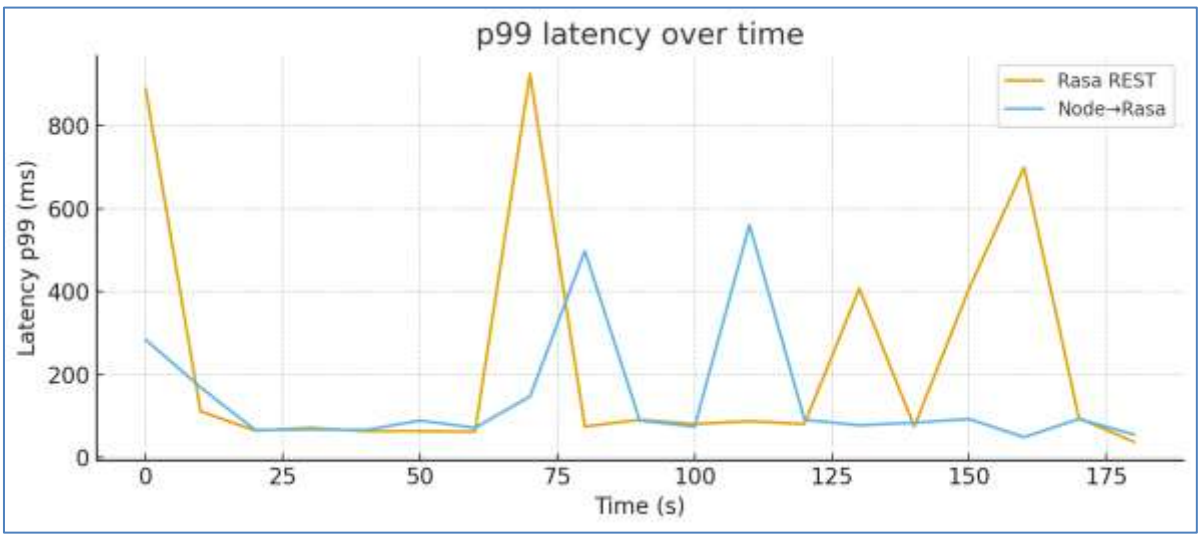
Tabel 2 Hasil *latency test* sistem

Komponen	Req	p50 (ms)	p95 (ms)	p99 (ms)	max (ms)	Error
Rasa REST	2700	32.1	89.1	584.2	974	0
Node → Rasa	2700	32.1	66.0	284.3	576	0

Tabel 2 merangkum hasil pengujian performa latensi. Pada *load* jalur *Node* menuju *Rasa* mencapai p50 32,1 ms, p95 66 ms, p99 284,3 ms dari 2700 *request* dengan 0 *error*, melampaui SLO yang telah ditetapkan. Dibanding *Rasa REST*, median hampir identik, sedangkan p99 lebih rendah, yang kami atribusikan pada penggunaan koneksi *keep-alive* pada *bridge*. Temuan ini menunjukkan *Node bridge* tidak menjadi *bottleneck* dan bahkan membantu menstabilkan *tail*.



Gambar 5 Grafik *request* per detik dan waktu



Gambar 6 Grafik latensi p99 dan waktu

Grafik RPS yang ditunjukkan pada Gambar 5 memperlihatkan transisi yang bersih dari 5rps ke 20rps sekitar detik ke-60 pada kedua jalur, indikasi *traffic generator* dan server *path* stabil. Gambar 6 menampilkan grafik p99, jalur *Rasa REST* menunjukkan beberapa lonjakan ratusan milidetik hingga mendekati 1 detik di sebagian *interval*, sedangkan *Node* menuju *Rasa* cenderung lebih rendah. Temuan ini konsisten dengan hipotesis bahwa *keep-alive* dan *connection pooling* pada *bridge* menekan beban *handshake* per-*request* sehingga *tail latency* lebih terkendali. Karena *LLM* tidak dilibatkan, hasil ini merefleksikan kapasitas jalur deterministik.

 ISSN:2302-8149
e-ISSN:2540-9719

Sebagai pelengkap, profil *action instrument* pada jalur *fallback LLM* menunjukkan median waktu panggilan berada di 13,29s dengan nilai puncak 14,51s selama pengujian. Durasi aksi action_llm_fallback praktis identik dengan waktu panggilan *LLM* sehingga hampir seluruh waktu eksekusi terserap oleh komputasi *LLM*, sementara aksi *non-LLM* berjalan pada satuan milidetik. Jejak memori proses saat *fallback* berada di angka 20–31 MiB dan tidak menunjukkan eskalasi yang mengkhawatirkan. Temuan ini memperkuat keputusan untuk memusatkan uji beban pada jalur deterministik serta menegaskan kebutuhan kebijakan pemanggilan *LLM* yang selektif melalui ambang batas yang tepat, penghematan konteks dan token, serta penerapan *circuit breaker* dan *timeout* agar biaya dan *tail latency* tetap terkendali di lingkungan *production*.

Temuan pada pengujian jalur deterministik dan profil jalur generatif menunjukkan bahwa pemisahan peran antara komponen deterministik dan generatif tidak hanya relevan untuk domain STIFIn, tetapi juga mencerminkan pola arsitektur yang dapat diterapkan pada domain lain dengan karakteristik serupa. Selama pengetahuan domain dapat direpresentasikan dalam bentuk *intent, entity*, dan aturan yang terkurasi, jalur deterministik berbasis *Rasa* dapat berfungsi sebagai jalur utama dengan latensi yang terprediksi, sementara *LLM* diposisikan sebagai mekanisme pelengkap. Stabilitas jalur deterministik serta tingginya *overhead* pada jalur generatif mengindikasikan bahwa efektivitas pendekatan ini bergantung pada kejelasan pembagian peran dan pemisahan evaluasi performa. Dengan penyesuaian ontologi domain dan ambang *fallback* berbasis *confidence NLU*, arsitektur ini dapat direplikasi pada konteks domain-spesifik lainnya.

5 Kesimpulan

Penelitian ini menunjukkan bahwa arsitektur *chatbot* hibrida untuk domain STIFIn dapat beroperasi secara stabil dan efisien pada jalur deterministik. Pengujian performa memperlihatkan bahwa baik akses *Rasa REST* maupun *Node.js bridge* memenuhi *service level objective (SLO)* dengan median latensi sekitar 32 ms tanpa *error*, serta menunjukkan bahwa penggunaan *bridge* tidak menimbulkan *overhead* signifikan dan bahkan membantu menekan *tail latency* melalui mekanisme *keep-alive* dan *connection pooling*. Uji fungsional menegaskan konsistensi alur *intent–entity–action* dengan penyimpanan konteks mk dan jk yang stabil, sementara *fallback* ke *LLM* hanya diaktifkan secara selektif ketika pertanyaan melampaui cakupan data latih *intent*. Profil eksekusi menunjukkan bahwa hampir seluruh latensi pada jalur *fallback* berasal dari komputasi *LLM*, yang menegaskan pentingnya pemisahan evaluasi performa dan *policy* pemanggilan *LLM* yang terkontrol. Pengembangan selanjutnya diarahkan pada perluasan pengetahuan deterministik, seperti peningkatan kualitas data pelatihan *intent* dan anotasi dari riwayat pesan masuk oleh pengguna, serta penerapan *guardrail* semantik dan *policy* pemanggilan *LLM* yang lebih ketat agar performa, efisiensi biaya, dan ketahanan terhadap halusinasi tetap terjaga pada skala yang lebih besar.

Referensi

[1] A. W. Paracha, U. Arshad, R. H. Ali, Z. Ul Abideen, M. H. Shah, T. Ali Khan, A. Z. Ijaz, N. Ali, A. B. Siddique, "*Leveraging AI and NLP in Chatbot Development: An Experimental Study*," in *2023 International Conference on Frontiers of Information Technology (FIT)*, Dec. 2023, pp. 172–177. DOI: 10.1109/FIT60620.2023.00040.

[2] M. Foosherian, H. Purwins, P. Rathnayake, T. Alam, R. Teimao, and K.-D. Thoben, "*Enhancing Pipeline-based Conversational Agents with Large Language Models*," Sept. 07, 2023, *arXiv*: arXiv:2309.03748. DOI: 10.48550/arXiv.2309.03748.

[3] S. V and S. Sampangi, "*Implementation of an Educational Chatbot using Rasa Framework*," *Int. J. Innov. Technol. Explor. Eng.*, Vol. 11, pp. 29–35, Aug. 2022, DOI: 10.35940/ijitee.G9189.0811922.

[4] K. G. Yager, "*Domain-Specific ChatBots for Science using Embeddings*," *Digit. Discov.*, Vol. 2, No. 6, pp. 1850–1861, 2023, DOI: 10.1039/D3DD00112A.

[5] R. Browne, M. M. Alam, Q. Saleem, A. Hyder, T. Kudo, F. D'Agresti, M. Maggio, K. Homma, E.-J. Siitonen, N. Kounosu, K. Jokinen, M. McTear, G. Napolitano, K. Kim, J. Tsujii, R. Wieching, T. Ogawa, and Y. Taki, "*Reflective Dialogues with a Humanoid Robot Integrated with an LLM and a Curated NLU System for Positive Behavioral Change in Older Adults*," *Electronics*, Vol. 13, No. 22, Art. No. 22, Jan. 2024, DOI: 10.3390/electronics13224364.

[6] W. Shalaby, A. Arantes, T. GonzalezDiaz, and C. Gupta, "*Building Chatbots from Large Scale Domain-Specific Knowledge bases: Challenges and Opportunities*," in *2020 IEEE International Conference on Prognostics and Health Management (ICPHM)*, Detroit, MI, USA: IEEE, June 2020, pp. 1–8. DOI: 10.1109/ICPHM49022.2020.9187036.

[7] D. Wulandari and J. S. Wibowo, "Implementasi *Chatbot* menggunakan *Framework* Rasa untuk Layanan Informasi Wisata di Kota Pati," *INTECOMS J. Inf. Technol. Comput. SCI.*, Vol. 6, No. 2, pp. 794–801, Sept. 2023, DOI: 10.31539/intecoms.v6i2.7107.

[8] V. S. Garófalo-Jerez, W. Hojas-Mazo, M. Moreno-Espino, Y. Villuendas-Rey, A. López-González, F. Maciá-Pérez, and J. V. Berná-Martínez, "*RESTful API for Intent Recognition based on RASA*," in *Advances in Soft Computing*, L. Martínez-Villaseñor and G. Ochoa-Ruiz, Eds., Cham: Springer Nature Switzerland, 2025, pp. 211–223. DOI: 10.1007/978-3-031-75543-9_16.

[9] L. Anindyati, "Analisis dan Perancangan Aplikasi *Chatbot* menggunakan *Framework* Rasa dan Sistem Informasi Pemeliharaan Aplikasi (Studi Kasus: Chatbot Penerimaan Mahasiswa Baru Politeknik Astra)," *J. Teknol. Inf. Dan Ilmu Komput.*, Vol. 10, No. 2, Art. No. 2, Apr. 2023, DOI: 10.25126/jtiik.20231026409.

[10] M. Joshi and R. K. Sharma, "*An Analytical Study and Review of Open Source Chatbot Framework, RASA*," *Int. J. Eng. Res. Technol.*, Vol. 9, No. 6, June 2020, DOI: 10.17577/IJERTV9IS060723.

[11] N. Darapaneni, A. R. Paduri, U. Tank, B. KanthaSamy, A. Ranjan, and R. Krisnakumar, "*Conversational AI: A Study on Capabilities and Limitations of Dialogue based System*," in *Multi-disciplinary Trends in Artificial Intelligence*, R. Morusupalli, T. S. Dandibhotla, V. V. Atluri, D. Windridge, P. Lingras, and V. R. Komati, Eds., Cham: Springer Nature Switzerland, 2023, pp. 503–512. DOI: 10.1007/978-3-031-36402-0_47.

[12] A. Liu *et al.*, "DeepSeek-V3 Technical Report," Feb. 18, 2025, *arXiv*: arXiv:2412.19437. DOI: 10.48550/arXiv.2412.19437.

[13] DeepSeek, "*Pricing Details (USD).*" Accessed: Oct. 11, 2025. [Online]. Available: https://api-docs.deepseek.com/quick_start/pricing-details-usd

[14] J. Jeong and J. Ahn, "*Accelerating LLM Serving for Multi-Turn Dialogues with Efficient Resource Management*," in *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, New York, NY, USA: Association for Computing Machinery, 2025, pp. 1–15. Accessed: Oct. 10, 2025. [Online]. Available: https://doi.org/10.1145/3676641.3716245

[15] M. R. A. Fajri and B. Hartono, "Pengembangan Aplikasi Chatbot Telegram menggunakan *Framework* Rasa untuk Pelayanan Administrasi di Perguruan Tinggi Universitas Stikubank," *J. JTIK J. Teknol. Inf. Dan Komun.*, Vol. 8, No. 1, Art. No. 1, Jan. 2024, DOI: 10.35870/jtik.v8i1.1370.

[16] A. D. Ferdian and S. N. Anwar, "Pengembangan *Chatbot* untuk Informasi Wisata Interaktif di Tangerang Selatan menggunakan *Framework* Rasa," *J. Teknol. Dan Sist. Inf. Bisnis*, Vol. 5, No. 4, Art. No. 4, Oct. 2023, DOI: 10.47233/jteksis.v5i4.953.

[17] Y. Windiatmoko, R. Rahmadi, and A. F. Hidayatullah, "*Developing Facebook Chatbot based on Deep Learning using RASA Framework for University Enquiries*," *IOP Conf. Ser. Mater. SCI. Eng.*, Vol. 1077, No. 1, p. 012060, Feb. 2021, DOI: 10.1088/1757-899X/1077/1/012060.

[18] A. Rachman, I. Mardhiyah, and M. Jannah, "Implementasi *Chatbot FAQ* pada Aplikasi Monev Kinerja Direktorat Jenderal Anggaran menggunakan *Framework Rasa Open Source*," *klik kaji. ilm. inform. dan Komput.*, Vol. 4, No. 1, Art. No. 1, Aug. 2023, DOI: 10.30865/klik.v4i1.1020.

[19] A. Pratiwi, M. P. Utami, and Y. Valentia, "*WhatsApp Story as a Medium of Digital Marketing Communication for MSME in Bogor Regency*," *AICCON*, Vol. 1, pp. 48–60, Feb. 2024.

[20] *"STIFIn Official Website*." Accessed: Oct. 10, 2025. [Online]. Available: https://stifin.com/

[21] A. Raza, M. Latif, M. U. Farooq, M. A. Baig, M. A. Akhtar, and Waseemullah, "*Enabling Context-based AI in Chatbots for Conveying Personalized Interdisciplinary Knowledge to Users*," *Eng. Technol. Appl. SCI. Res.*, Vol. 13, No. 6, Art. No. 6, Dec. 2023, DOI: 10.48084/etasr.6313.