

Implementasi Arsitektur *Self-Reflective* RAG (*Gemma 4*) pada *Edge Computing* untuk Layanan Psychological First Aid dan Edukasi PTM di Dinas Kesehatan Kota Semarang

Laporan Kegiatan Magang

Oleh

BINSARNIARI SINAGA 2602184012



Research Track

Digital Business

BINUS BUSINESS SCHOOL

BINA NUSANTARA UNIVERSITY

SEMARANG

2026

Implementasi Arsitektur *Self-Reflective* RAG (*Gemma 4*) pada *Edge Computing* untuk Layanan Psychological First Aid dan Edukasi PTM di Dinas Kesehatan Kota Semarang

Laporan Kegiatan Magang

**diajukan sebagai salah satu syarat
untuk penyelesaian kegiatan magang**

Oleh

BINSARNIARI SINAGA 2602184012



Research Track

DIGITAL BUSINESS

BINUS BUSINESS SCHOOL

BINA NUSANTARA UNIVERSITY

SEMARANG

2026

LEMBAR PENGESAHAN

LEMBAR PENGESAHAN

Laporan akhir kegiatan magang riset di Dinas Kesehatan Kota Semarang dengan judul:

**IMPLEMENTASI ARSITEKTUR SELF-REFLECTIVE RAG (GEMMA 4) PADA EDGE COMPUTING
UNTUK LAYANAN PSYCHOLOGICAL FIRST AID DAN EDUKASI PTM DI DINAS KESEHATAN KOTA
SEMARANG**

Disusun oleh :

Nama : Binsarniari Sinaga

NIM : 2602184012

Program Studi : Digital Business

Telah disahkan pada :

Hari : Rabu

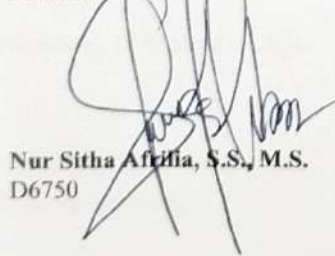
Tanggal : 30 September 2026

Dosen Pembimbing



Prof. Dr. Dra. Ngatindriatun, M.P.
D5311

Mengetahui
Koordinator BINUS Research and Technology
Transfer



Nur Sitha Affilia, S.S., M.S.
D6750

Sub Koordinator Informasi dan Pengendalian
Sarana Kesehatan Dinas Kesehatan Kota
Semarang



Hanif Pandu Suhito, SKM, M.Kom, M.Si.
NIP: 198402192005011003

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan Laporan Akhir Enrichment Program jalur *Research Track* dengan judul "**Implementasi Arsitektur Self-Reflective RAG (Gemma 4) pada Edge Computing untuk Layanan Psychological First Aid dan Edukasi PTM di Dinas Kesehatan Kota Semarang**".

Laporan ini disusun sebagai salah satu syarat penyelesaian mata kuliah Enrichment Program di Program Studi Digital Business, BINUS University. Penulis menyadari bahwa penyelesaian riset ini melalui serangkaian proses teknis dan akademis yang panjang. Berkat arahan, bimbingan, dan dukungan dari berbagai pihak, penelitian ini dapat diselesaikan dengan hasil yang diharapkan.

Pada kesempatan ini, penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada:

1. **Dicky Jhon Anderson Butarbutar, S.E., S.Kom., M.M.**, selaku Koordinator *Enrichment Program* yang telah memfasilitasi pelaksanaan program *Research Track*.
2. **Ibu Nur Sitha Afrilia, S.S., M.S.**, selaku Koordinator *Research and Technology Transfer* yang telah membantu memfasilitasi magang riset di Dinas Kesehatan Kota Semarang
3. **Prof. Dr. Dra. Ngatindriatun, M.P.**, selaku *Faculty Supervisor* yang dengan penuh kesabaran telah memberikan bimbingan, arahan, dan wawasan ilmiah kepada penulis.
4. **Bapak Dr. dr. Mochamad Abdul Hakam, Sp.PD.**, selaku pimpinan Dinas Kesehatan Kota Semarang
5. **Bapak Hanif Pandu Suhito, SKM, M.Kom., M.Si., Sholikun, S.Inf.Med., dan Ibu Prahita Indriana Rantiasmi, SKM.**, selaku pembimbing lapangan di Dinas Kesehatan Kota Semarang yang selalu sabar memberikan arahan teknis, wawasan dunia kerja, serta ilmu baru selama pengembangan proyek Implementasi Arsitektur Self-Reflective RAG.
6. **seluruh staf dan pegawai di lingkungan Dinas Kesehatan Kota Semarang**, yang telah menerima kehadiran penulis dengan ramah.
7. **seluruh dosen program studi Digital Business BINUS University**, yang telah membekali penulis dengan fondasi ilmu bisnis digital, manajemen pengetahuan, dan analitik data.
8. **orang tua dan keluarga**, yang senantiasa memberikan doa, dukungan moral, serta dorongan semangat tanpa henti.
9. Semua pihak yang tidak dapat penulis sebutkan, yang telah membantu kelancaran penulisan laporan ini

Penulis menyadari bahwa laporan ini masih memiliki keterbatasan. Oleh karena itu, kritik dan saran yang membangun sangat penulis harapkan demi penyempurnaan karya di masa mendatang. Semoga laporan penelitian ini dapat memberikan kontribusi nyata bagi pengembangan teknologi *Edge AI*, transformasi digital sektor kesehatan publik, serta peningkatan kualitas pelayanan masyarakat di Kota Semarang.

Semarang, 30 September 2026

Binsarniari Sinaga

DAFTAR ISI

LEMBAR PENGESAHAN.....	3
KATA PENGANTAR.....	4
DAFTAR ISI.....	5
BAB 1: COMPANY PROFILE.....	7
1.1 Profil Perusahaan/Institusi.....	7
1.2 Profil dan Peran Mahasiswa.....	8
BAB 2:RESEARCH REPORT.....	10
2.1 Laporan Penelitian.....	10
2.1.1. Abstract.....	10
2.1.2 Pendahuluan.....	11
2.1.2.1 Latar Belakang Penelitian.....	11
2.1.2.2 Kesenjangan Penelitian (Research Gap).....	11
2.1.2.3 Tujuan dan Kontribusi Penelitian.....	11
2.1.2.4 Research Question.....	11
2.1.2.5 Research Hypothesis.....	12
2.1.2.6 Research Limitations.....	13
2.1.2.7 Benefit of Research.....	13
2.1.3 Literature Review.....	13
2.1.3.1 Paradigma Kesehatan Urban dan Transformasi SIKDA Kota Semarang.....	13
2.1.3.2 Evolusi Large Language Model: Kemampuan Arsitektural dan Kognitif Gemma 4.....	14
2.1.3.3 Optimasi Edge Computing Terdistribusi dan Ketahanan Privasi Data.....	15
2.1.3.4 Evaluasi Fungsionalitas Sistem Melalui Black Box Testing.....	16
2.1.3.5 Evaluasi Pengalaman Pengguna dengan System Usability Scale (SUS).....	16
2.1.3.6 Technology Acceptance Model (TAM) dalam Adopsi AI Kesehatan.....	17
2.1.3.7 Antropomorfisme dan Social Presence pada Conversational Agent.....	17
2.1.4 Research Methodology.....	18
2.1.4.1 Desain Penelitian.....	18
2.1.4.2 Populasi dan Sampel.....	18
2.1.4.3 Definisi Operasional dan Variabel Penelitian.....	19
2.1.4.4 Teknik Pengumpulan Data.....	21
2.1.4.5 Teknik Analisis Data.....	23
2.1.4.5.1 Analisis Kuantitatif Teknis (Performance & RAGAS Benchmarking).....	23
2.1.4.5.2 Analisis Kuantitatif UX & Penerimaan Sistem (SUS & Two-Stage PLS-SEM).....	24
2.1.4.5.3 Analisis Kualitatif (Thematic Analysis).....	25
2.1.5 Research Results.....	26
2.1.5.1 Deskripsi Artefak Teknologi & Lingkungan Edge Computing.....	26
2.1.5.2 Evaluasi Kinerja Teknis Artefak (Benchmarking Latensi & Akurasi N=115).....	27
2.1.5.3 Evaluasi Latensi Response Komputasi Edge.....	27
2.1.5.4 Evaluasi Akurasi Jawaban Berdasarkan Kategori Kueri.....	28
2.1.5.5 Evaluasi Fungsionalitas & Keselamatan Sistem (Black Box Testing).....	30

2.1.5.6 Evaluasi Model Pengukuran / Outer Model PLS-SEM (Stage 1).....	30
2.1.5.7 Evaluasi Model Struktural / Inner Model & HOC (Stage 2).....	32
2.1.5.7.a Outer Weights & Loadings Pembentukan HOC.....	32
2.1.5.7.b Koefisien Determinasi (R ²) dan Ukuran Efek (f ²).....	33
2.1.5.7.c Pengujian Hipotesis Pengaruh Langsung dan Mediasi.....	34
2.1.5.8 Evaluasi Usability (SUS) & Analisis Kualitatif.....	36
2.1.5.8.a Perhitungan System Usability Scale (SUS).....	36
2.1.5.8.b Thematic Analysis Feedback Kualitatif.....	36
2.1.5.9 Pembahasan Integratif DSR & Temuan Grounding.....	36
2.1.6 Conclusions.....	37
2.1.6.1 Kesimpulan.....	37
2.1.6.2 Implikasi Penelitian.....	38
2.1.6.3 Keterbatasan dan Saran Penelitian Selanjutnya (Further Research).....	38
2.2 Kegiatan yang Dilakukan Sesuai Learning Plan/Project.....	39
A. Identifikasi Masalah Teknis: Tingkat halusinasi dan masalah koneksi (Minggu Kedua Bulan ke-3).....	40
B. Pemecahan Masalah: Perubahan Model ke Gemma 4 (Pertengahan Bulan ke-3).....	41
C. Penuntasan Kendala Teknis Analisis (Bulan ke-5 & ke-6).....	41
BAB 3: KESIMPULAN.....	42
3.1 Kesimpulan Pelaksanaan Program.....	42
3.2 Saran.....	43
REFERENCES.....	44
LAMPIRAN.....	47

BAB 1: COMPANY PROFILE

1.1 Profil Perusahaan/Institusi

Nama Institusi: Dinas Kesehatan Kota Semarang

Unit Kerja: Bidang SDK – Informasi Kesehatan (INFOKES)



Sekilas Tentang Perusahaan:

Visi: *“Kota Semarang Menjadi Pusat Ekonomi yang Maju, Berkeadilan Sosial, Lestari dan Inklusif”*

Misi:

- Mewujudkan pemerataan pendidikan dan kesejahteraan sosial masyarakat yang toleran dan berbudaya dalam semangat kebhinekaan, serta meningkatkan pembangunan manusia yang produktif, berkualitas dan berkepribadian.
- Mewujudkan kesehatan seluruh masyarakat yang berfokus pada kebutuhan individu dengan mengutamakan aspek pencegahan, pengobatan, dan rehabilitasi.
- Mewujudkan pemenuhan kebutuhan dasar berupa ketersediaan pangan, sandang, dan papan yang merupakan elemen kunci agar tercapai taraf hidup yang layak.
- Mewujudkan perekonomian inklusif melalui penyediaan lapangan kerja dengan membangun kemandirian ekonomi kerakyatan berbasis potensi sumber daya lokal, dan peningkatan daya saing sumber daya manusia dengan pemanfaatan teknologi digital.
- Mewujudkan infrastruktur kota yang saling terhubung dengan peningkatan aksesibilitas dan konektivitas wilayah yang berkelanjutan.
- Mewujudkan kualitas lingkungan kota yang tangguh, berkelanjutan, sekaligus meningkatkan pengendalian banjir, rob, serta dampaknya bagi masyarakat.
- Mewujudkan pelayanan publik dan tata kelola pemerintahan yang berkualitas, dinamis, bersih, bebas dari korupsi, berkeadaban, dan inklusif berbasis kota cerdas.

1.2 Profil dan Peran Mahasiswa

Posisi: *Research Intern* (Penelitian Magang) **Unit Penempatan:** *Research Track - Digital Business*

Deskripsi Peran: Dalam skema *Enrichment Program* jalur Penelitian (*Research Fellowship Track*), mahasiswa tidak sekadar berperan sebagai peserta didik, melainkan sebagai peneliti mandiri yang bekerja di bawah supervisi *Site Supervisor* dan *Faculty Supervisor*. Peran ini menuntut kedisiplinan tinggi, kemampuan analisis kritis, dan manajemen proyek riset yang terstruktur.

Tanggung Jawab Utama: Sebagai *Research Intern*, tanggung jawab utama penulis meliputi:

1. **Perancangan Desain Riset & Arsitektur Sistem (*Research & System Design*):** Mengidentifikasi kesenjangan penelitian (*research gap*) terkait adopsi kecerdasan buatan terdistribusi (*Edge AI*) pada sektor kesehatan publik, serta merancang kerangka kerja *Design Science Research (DSR)* untuk pengembangan asisten digital "Emi".
2. **Akuisisi & Pemrosesan Data Kesehatan (*Data Ingestion & Curation*):** Mengumpulkan, membersihkan, dan mengonversi korpus dokumen resmi Dinas Kesehatan Kota Semarang (Profil Kesehatan 2024, Peta Risiko, dan Laporan *UNICEF* 2024) menggunakan mesin *GOT-OCR 2.0* ke dalam *Qdrant Vector Database* untuk menopang pipa *Retrieval-Augmented Generation (RAG)*.
3. **Pengembangan & Optimasi Artefak AI (*AI Artifact Engineering*):** Menerapkan kompresi model *Google Gemma 4 12B* menggunakan teknik *Quantization-Aware Training (QAT 4-bit GGUF)* pada *Edge GPU* lokal (*VRAM ≤ 16GB Dual-GPU*), merancang *Affective System Prompting*, serta mengeksekusi pengujian fungsionalitas (*Black Box Testing*) dan *benchmarking* latensi/*RAGAS* ($N = 115$).
4. **Evaluasi Pengguna & Analisis Statistik (*User Evaluation & Analytics*):** Menyelenggarakan *User Acceptance Testing (UAT)* pada Generasi Z Kota Semarang ($N = 20$), mengolah data survei menggunakan pendekatan *Two-Stage Higher-Order Construct (HOC) PLS-SEM* pada *SmartPLS 4*, mengkalkulasi *System Usability Scale (SUS)*, serta menganalisis umpan balik kualitatif ($n = 11$) melalui *Thematic Analysis*.

5. **Penulisan Laporan Akhir & Publikasi Ilmiah (*Scientific Writing & Reporting*):** Menyusun manuskrip laporan penelitian utuh (Bab 1 sampai Bab 5), merumuskan abstrak bilingual, serta mensintesis temuan kinerja komputasi *edge* dengan teori *Human-Computer Interaction* (HCI) sesuai standar publikasi ilmiah.

BAB 2:RESEARCH REPORT

2.1 Laporan Penelitian

2.1.1. Abstract

Tantangan Penyakit Tidak Menular (PTM) dan kesehatan mental di kalangan Generasi Z Kota Semarang memerlukan tersedianya sistem informasi kesehatan yang interaktif, empatik, serta menjamin privasi data. Namun, penerapan kecerdasan buatan (AI) tingkat lanjut di sektor publik kerap terkendala oleh tingginya biaya infrastruktur komputasi awan (*cloud*). Penelitian ini bertujuan untuk merancang, mengimplementasikan, dan mengevaluasi asisten digital "Emi" berbasis arsitektur *Self-Reflective Retrieval-Augmented Generation (RAG)* menggunakan model *Google Gemma 4 12B* terkuantisasi 4-bit (GGUF) pada perangkat keras komputasi lokal (*Edge Computing* dengan $VRAM \leq 16\text{ GB}$). Mengadopsi metodologi *Design Science Research (DSR)*, evaluasi komprehensif dilakukan melalui pengujian kinerja teknis (N=115 rekaman interaksi), uji fungsionalitas (*Black Box Testing*), serta uji penerimaan pengguna (N=20 responden Generasi Z) menggunakan pemodelan *Two-Stage Higher-Order Construct (HOC) PLS-SEM* dan *System Usability Scale (SUS)*. Hasil pengujian menunjukkan rata-rata akurasi keseluruhan sistem mencapai 70,22% dengan rata-rata latensi sebesar 39,66 detik. Secara khusus, akurasi untuk intervensi krisis *Psychological First Aid (PFA)* mencapai 93,75% dan dukungan emosional sebesar 80,77%. Pengujian fungsionalitas melalui *Black Box Testing* menunjukkan tingkat kelulusan 100%. Hasil analisis PLS-SEM mengonfirmasi bahwa niat perilaku adopsi pengguna secara signifikan dipengaruhi oleh konstruk *Social-Trust* ($\beta = 0,981$, $p = 0,000$, $R^2 = 83,3\%$), sementara konstruk *Utilitarian* tidak berpengaruh secara signifikan ($p = 0,319$). Skor rata-rata SUS yang diperoleh adalah sebesar 59,63 (*Grade D / Marginal Low*), yang menunjukkan adanya kesenjangan adopsi (*Adoption Gap*) akibat tingginya durasi tunggu respons pada kueri medis yang kompleks (rata-rata = 67,90 detik). Hasil penelitian ini menyajikan cetak biru penerapan Edge AI yang efisien dari segi anggaran dan menjaga privasi bagi instansi pemerintah daerah.

Kata Kunci: *Edge Computing, Gemma 4, Psychological First Aid, Retrieval-Augmented Generation, Two-Stage PLS-SEM.*

2.1.2 Pendahuluan

2.1.2.1 Latar Belakang Penelitian

Kesehatan masyarakat urban di Kota Semarang saat ini menghadapi tantangan serius akibat adanya transisi epidemiologi dari penyakit menular ke Penyakit Tidak Menular (PTM). Berdasarkan Profil Kesehatan Kota Semarang Tahun 2024 yang dirilis secara resmi oleh Dinas Kesehatan, tercatat 288.750 penderita hipertensi yang telah menerima pelayanan kesehatan tingkat dasar. Lebih lanjut, dari total 1.114.865 penduduk usia produktif (15-59 tahun) yang menjalani skrining kesehatan, ditemukan sebanyak 26,4% populasi perempuan dan 21,8% populasi laki-laki berada dalam kategori berisiko. Di samping itu, prevalensi kasus diabetes mellitus di Semarang tercatat mencapai 34.561 jiwa pada periode pelaporan yang sama (Dinas Kesehatan Kota Semarang, 2024b).

Pada domain kesehatan mental, situasinya memerlukan perhatian yang tidak kalah serius. Berdasarkan laporan *nasional Indonesia Adolescent Health Profile 2024* oleh UNICEF, 1 dari 3 remaja (setara dengan sekitar 15,5 juta jiwa) menghadapi tantangan kesehatan mental, namun hanya sekitar 10,4% yang mencari pengobatan atau bantuan profesional (UNICEF Indonesia, 2024). Di tingkat lokal Kota Semarang, tercatat terdapat 3.872 kasus Orang dengan Gangguan Jiwa (ODGJ) berat pada kelompok usia produktif. Hambatan utama dalam penanganan krisis kesehatan ini bukan sekadar terbatas pada ketersediaan fisik layanan medis, melainkan rendahnya aksesibilitas informasi yang sering kali terhalang oleh stigma sosial serta kesulitan akses sistem informasi kesehatan publik.

2.1.2.2 Kesenjangan Penelitian (*Research Gap*)

Meskipun data kesehatan publik tersedia secara melimpah dalam format dokumen laporan statis (Boateng et al., 2024), terdapat kesenjangan besar dalam hal pencarian informasi (*knowledge retrieval*) yang interaktif bagi masyarakat awam. Sistem informasi kesehatan daerah yang ada saat ini masih bersifat searah dan sulit diakses secara cepat menggunakan bahasa alami sehari-hari. Di sisi teknologi, implementasi model kecerdasan buatan (AI) canggih sering kali terbentur oleh tingginya biaya operasional server berbasis awan (Zheng et al., 2025). Masih terbatas penelitian yang mengoptimasi sistem *Retrieval-Augmented Generation (RAG)* pada infrastruktur komputasi dengan spesifikasi terbatas (*low-resource hardware*) setingkat komputer kantor untuk menjembatani data statistik formal dengan kebutuhan informasi kesehatan harian masyarakat (Qu et al., 2025).

2.1.2.3 Tujuan dan Kontribusi Penelitian

Riset ini bertujuan mengembangkan sistem Digital Health Assistant yang mampu melakukan ekstraksi informasi dari dokumen resmi Dinkes Semarang secara otomatis. Kontribusi utamanya adalah menyediakan skema implementasi model AI (seperti *Gemma 4*) (Gemma Team, 2026) yang ringan namun akurat untuk memberikan edukasi pencegahan PTM dan deteksi dini kesehatan mental tanpa memerlukan investasi perangkat keras kelas tinggi.

2.1.2.4 *Research Question*

Berdasarkan kesenjangan penelitian yang teridentifikasi, dirumuskan tiga pertanyaan penelitian utama (*Research Questions/RQ*) sebagai berikut:

1. **RQ1:** Bagaimana tingkat ketepatan faktual (*faithfulness*) dan relevansi respons (*answer relevancy*) sistem *Self-Reflective Retrieval-Augmented Generation (RAG)* berbasis model *Gemma 4* dalam mengasimilasi dataset Profil Kesehatan Kota Semarang Tahun 2024 untuk kueri edukasi kesehatan masyarakat?
2. **RQ2:** Sejauh mana implementasi teknik kuantisasi *low-bit* (4-bit GGUF via QAT) pada model bahasa berparameter besar (Google Gemma 4 12B) di lingkungan *edge computing* lokal dapat mempertahankan efektivitas penalaran (*Chain-of-Thought*) medis asisten digital tanpa mengalami degradasi fungsional yang drastis?
3. **RQ3:** Bagaimana pengaruh konstruk sosial-afektif (*Social-Trust*) dan utilitas teknis (*Utilitarian*) terhadap niat perilaku pengguna Generasi Z dalam mengadopsi asisten digital kesehatan ini berdasarkan evaluasi pemodelan struktural PLS-SEM?

2.1.2.5 Research Hypothesis

- **H1: Efektivitas Akurasi Pengetahuan**

Implementasi mekanisme *Retrieval-Augmented Generation (RAG)* menggunakan dokumen primer resmi Dinas Kesehatan Kota Semarang secara signifikan akan meningkatkan ketepatan faktual (*faithfulness*) serta meminimalkan risiko halusinasi informasi pada model *Gemma 4* dibandingkan dengan model tanpa integrasi data eksternal (Es et al., 2023).

- **H2: Efisiensi Komputasi Edge**

Penerapan teknik kuantisasi 4-bit (format GGUF) pada model *Google Gemma 4 12B* memungkinkan sistem asisten kesehatan berjalan secara efisien pada perangkat keras lokal (*edge GPU* kelas konsumen) tanpa mengakibatkan penurunan kemampuan penalaran (*reasoning*) medis yang drastis (Egiazarian et al., 2024).

- **H3: Penerimaan Sistem Secara Afektif dan Usability**

Personalisasi respons melalui instruksi sistem berbasis empati (*Affective Prompting*) pada model *Gemma 4 12B* akan meningkatkan kenyamanan interaksi pengguna Generasi Z, yang secara struktural diukur melalui integrasi model penerimaan teknologi (TAM) dan model persamaan struktural (PLS-SEM) sebagai berikut:

- **H3a (HOC_SocTrust -> BI):** Konstruk *Social-Trust* (gabungan dari dimensi antropomorfisme, kehadiran sosial, dan tingkat kepercayaan) berpengaruh positif dan signifikan terhadap niat perilaku (*Behavioral Intention*) penggunaan sistem (Li et al., 2023).
- **H3b (HOC_Util -> BI):** Konstruk *Utilitarian* (gabungan dari dimensi persepsi kemudahan penggunaan dan persepsi kebermanfaatan) berpengaruh positif dan signifikan terhadap niat perilaku (*Behavioral Intention*) penggunaan sistem (Davis et al., 1989).
- **H3c (BI -> SUS):** Niat perilaku (*Behavioral Intention*) berpengaruh positif dan signifikan terhadap skor kepraktisan sistem yang dirasakan (*System Usability Scale/SUS*) (Brooke, 1996).
- **H3d (Mediasi):** Niat perilaku (*Behavioral Intention*) memediasi pengaruh konstruk *Social-Trust* terhadap skor kepraktisan sistem (SUS).

2.1.2.6 Research Limitations

Untuk memastikan fokus, validitas, dan kelayakan analisis, penelitian ini menetapkan batasan metodologis sebagai berikut:

- **Cakupan Pengetahuan Medis:** Korpus pengetahuan asisten digital dibatasi secara ketat pada dokumen resmi terkait Penyakit Tidak Menular (PTM) dari Profil Kesehatan Kota Semarang Tahun 2024, peta risiko daerah, serta panduan kesehatan mental remaja dari *UNICEF* Indonesia Tahun 2024. Asisten digital ini berfungsi sebagai media edukasi dan skrining awal, bukan sebagai pengganti diagnosis klinis psikolog atau dokter profesional.
- **Keterbatasan Perangkat Keras (*Edge GPU*):** Pengujian dan operasional sistem dijalankan secara lokal (on-premises) menggunakan kartu grafis kelas konsumen dengan keterbatasan memori video (VRAM ≤ 8 GB) untuk mensimulasikan keterbatasan infrastruktur komputer publik di instansi daerah.
- **Karakteristik dan Ukuran Sampel:** Evaluasi pengalaman pengguna (*User Acceptance Testing*) dibatasi pada N=20 responden Generasi Z (rentang usia 14-29 tahun) di Kota Semarang. Sementara itu, evaluasi kuantitatif terhadap performa dan ketepatan teknis dijalankan secara objektif pada N = 115 rekaman log interaksi sistem.

2.1.2.7 Benefit of Research

Penelitian ini diharapkan memberikan manfaat dan kontribusi yang signifikan dari berbagai aspek:

- **Digitalisasi Sektor Publik yang Efisien:** Menyajikan model praktis bagi instansi pemerintah daerah dalam mengeksekusi transformasi digital yang efisien secara anggaran, tanpa harus melakukan investasi server awan berbayar berbiaya tinggi.
- **Peningkatan Literasi Kesehatan Masyarakat:** Menyediakan media pencarian informasi kesehatan interaktif yang dekat dengan gaya hidup digital Generasi Z, guna memperluas jangkauan edukasi promotif dan preventif kesehatan daerah.
- **Pengembangan Keilmuan Bisnis Digital:** Memperkuat kerangka akademik bidang manajemen pengetahuan (*Knowledge Management Systems*) berbasis kecerdasan buatan, khususnya integrasi metodologi *Design Science Research (DSR)* dengan evaluasi *Human-Computer Interaction (HCI)*.

2.1.3 Literature Review

2.1.3.1 Paradigma Kesehatan Urban dan Transformasi SIKDA Kota Semarang

Kondisi kesehatan masyarakat urban di Kota Semarang saat ini menghadapi tantangan serius yang diakibatkan oleh transisi epidemiologi yang bergeser secara signifikan dari penyakit menular menuju Penyakit Tidak Menular (PTM). Berdasarkan data Profil Kesehatan Kota Semarang tahun 2024, beban PTM di wilayah ini tercatat cukup tinggi, dengan tercatatnya 288.750 penderita hipertensi yang telah mendapatkan akses pelayanan kesehatan tingkat dasar maupun lanjutan. Situasi ini tergolong signifikan berdasarkan hasil skrining terhadap 1.114.865 warga usia produktif (rentang usia 15-59 tahun), yang menemukan bahwa 26,4% populasi perempuan dan 21,8% populasi laki-laki berada dalam kategori berisiko tinggi terhadap PTM, diiringi dengan

temuan 34.561 penderita diabetes melitus pada periode pencatatan yang sama (Dinas Kesehatan Kota Semarang, 2024b).

Pada domain kesehatan mental, situasinya memerlukan perhatian yang tidak kalah serius. Berdasarkan laporan nasional *Indonesia Adolescent Health Profile 2024* oleh UNICEF, 1 dari 3 remaja (setara dengan sekitar 15,5 juta jiwa) menghadapi tantangan kesehatan mental, namun hanya sekitar 10,4% yang mencari pengobatan atau bantuan profesional (UNICEF Indonesia, 2024). Di tingkat lokal Kota Semarang, terdapat 3.872 orang dengan gangguan jiwa berat yang berada dalam kelompok usia produktif. Hambatan utama dalam menghadapi krisis kesehatan ini tidak hanya terletak pada kurangnya layanan medis yang tersedia, tetapi juga karena informasi yang kurang mudah diakses. Masalah ini seringkali terjadi karena adanya stigma sosial dan sistem informasi kesehatan umum yang masih kaku serta hanya mengalir satu arah.

Dalam konteks regulasi dan manajemen data, efektivitas Sistem Informasi Kesehatan Daerah (SIKDA) di tingkat kota memerlukan transformasi digital yang tidak hanya berfokus pada pencatatan administratif internal, tetapi juga perluasan akses informasi kesehatan publik yang inklusif, anonim, dan *non-judgmental*. Evaluasi empiris terhadap SIKDA di Provinsi Jawa Tengah oleh Lestari et al. (2016) menunjukkan bahwa dari tujuh komponen utama Sistem Informasi Kesehatan (SIK), empat komponen penting, yaitu pengelolaan (54,7%), sumber daya (54,2%), sumber data (58%), dan manajemen data (41,3%) dinilai "ada tetapi tidak adekuat". Lemahnya regulasi di tingkat daerah serta keterbatasan kapasitas tata kelola data mengakibatkan posisi pengelola SIK kurang optimal, sehingga memicu inkonsistensi prosedur pengumpulan data antara dinas kesehatan dan fasilitas kesehatan primer. Oleh karena itu, penyediaan asisten digital berbasis kecerdasan buatan (*conversational AI*) yang terjamin anonimitasnya serta bersifat *non-judgmental* bertindak sebagai *supplementary conversational layer* di atas SIKDA untuk memfasilitasi pencarian pengetahuan (*knowledge retrieval*) secara mandiri.

Dalam merancang asisten ini, interaksi penyampaian informasi medis wajib dilandasi oleh Teori Beban Kognitif (*Cognitive Load Theory*) dari Sweller (2011) guna memastikan bahwa instruksi dan saran yang diberikan sangat mudah dipahami, diakses secara intuitif, dan tidak membebani kapasitas memori kerja (*working memory*) pengguna. Lebih jauh lagi, pendekatan Computing Afektif (*Affective Computing*) murni dari Picard (1997, 2003) diintegrasikan agar sistem kecerdasan buatan tidak hanya sekadar menyajikan data mentah, melainkan secara aktif mampu merespons emosi pengguna melalui pola komunikasi yang sarat akan empati, sehingga berhasil membangun koneksi emosional yang menenangkan bagi pengguna yang sedang mengalami fase kecemasan.

2.1.3.2 Evolusi Large Language Model: Kemampuan Arsitektural dan Kognitif Gemma 4

Generasi terbaru dari keluarga model *Google*, yakni *Gemma 4*, menetapkan standar baru dalam efisiensi komputasi dan kemampuan penalaran melalui arsitektur multimodal yang secara bawaan bersifat efisien. Model ini dirancang dalam berbagai ukuran, termasuk arsitektur *dense* (seperti varian 12B) dan *Mixture-of-Experts* (MoE) (Gemma Team, 2026). Dalam menangani intervensi krisis kesehatan jiwa (*Psychological First Aid/PFA*) dan edukasi Penyakit Tidak Menular (PTM) yang membutuhkan empati dan penalaran tingkat tinggi, model *Gemma 4* sangat relevan karena menghadirkan inovasi *thinking mode*. Melalui *thinking mode* ini, model dapat

menghasilkan jejak penalaran (*reasoning traces*) terlebih dahulu sebelum memberikan respons akhir, yang terbukti meningkatkan kapabilitas model pada domain yang membutuhkan penalaran kompleks.

Secara arsitektural, implementasi Gemma 4 varian 12B memperkenalkan paradigma *encoder-free* yang terpadu. Arsitektur ini menggantikan *encoder* audio dan visi yang terpisah dengan modul proyeksi ringan, sehingga memungkinkan sistem memetakan potongan audio mentah berdurasi 40 ms dan *patch* gambar langsung ke dalam ruang *embedding* LLM. Pemrosesan dokumen ekstensif dalam RAG sering kali memicu ledakan memori pada *KV cache*. Untuk menjaga efisiensi pada perangkat dengan komputasi terbatas, Gemma 4 menerapkan rasio 5:1 untuk *sliding window* lokal terhadap *self-attention* global, penggunaan *positional encoding p-ROPE*, serta teknik *KV cache sharing* yang mampu mengurangi jejak memori *global KV cache* hingga 37,5% (Gemma Team, 2026).

Kendala keterbatasan perangkat keras pada fasilitas instansi pemerintah juga diatasi melalui implementasi kompresi model. Namun, terdapat perbedaan metodologis yang sangat krusial antara:

1. Quantization-Aware Training (QAT): Proses kuantisasi yang disimulasikan selama fase pelatihan atau fine-tuning model. Metode ini melatih bobot model untuk beradaptasi dengan keterbatasan representasi bit rendah, sehingga meminimalkan degradasi akurasi dan memelihara kemampuan penalaran berantai (*Chain-of-Thought/CoT*) secara optimal (Egiazarian et al., 2024).

2. Post-Training Quantization (PTQ) biasa pada format GGUF 4-bit: Kuantisasi langsung setelah model selesai dilatih tanpa proses latihan ulang tambahan. Meskipun PTQ lebih cepat dan praktis untuk diimplementasikan, metode ini rentan terhadap degradasi perplexity penalaran medis yang lebih besar pada tingkat kuantisasi bit rendah (*low-bit quantization*).

2.1.3.3 Optimasi Edge Computing Terdistribusi dan Ketahanan Privasi Data

Hambatan paling praktis dan utama dalam adopsi kecerdasan buatan tingkat lanjut di fasilitas instansi pemerintah daerah adalah keterbatasan ekstrem pada spesifikasi perangkat keras (*hardware*). Menurut analisis literatur oleh Yurez et al. (2026), transformasi digital di fasilitas kesehatan primer Indonesia menghadapi tantangan struktural yang berat, di mana lebih dari 80% fasilitas belum mengadopsi teknologi digital secara optimal. Lebih lanjut, survei nasional menunjukkan adanya kesenjangan infrastruktur yang nyata, dengan 21% puskesmas memiliki akses internet terbatas atau tidak tersedia sama sekali, serta 8,02% puskesmas beroperasi tanpa pasokan listrik 24 jam (Aisyah, Setiawan, Lokopessy, et al., 2025).

Dalam realitas infrastruktur yang terfragmentasi ini, implementasi model secara lokal melalui *komputasi edge* (*Edge Computing*) menjadi sangat relevan. Dengan menerapkan teknik kuantisasi 4-bit (format GGUF) hasil optimasi QAT pada model *Google Gemma 4 12B*, model berukuran besar ini dapat dikompresi hingga berukuran ≈ 10 GB, memungkinkannya untuk berjalan secara responsif pada kartu grafis kelas konsumen dengan memori video terbatas ($\text{VRAM} \leq 16$ GB) untuk mensimulasikan keterbatasan komputer instansi pemerintah di daerah. Pemanfaatan teknologi percepatan inferensi seperti PowerInfer (Song et al., 2024) memungkinkan perangkat keras lokal menangani beban permintaan dengan tingkat latensi yang terkontrol.

Eksekusi arsitektur model secara lokal pada perangkat edge ini memberikan jaminan keamanan privasi data yang tinggi melalui isolasi pemrosesan lokal (*on-device isolation*). Isolasi pemrosesan data di dalam ekosistem tertutup Dinas Kesehatan memastikan bahwa seluruh data percakapan sensitif dan data riwayat skrining kesehatan warganya ditangani seratus persen secara *on-premise*. Metode penanganan data lokal ini meniadakan transmisi data ke server pihak ketiga di luar negeri, sehingga memenuhi asas operasional *Privacy Preserving* (Edemacu & Wu, 2025) yang sejalan dengan kepatuhan hukum tertinggi dalam manajemen data sektor kesehatan publik.

2.1.3.4 Evaluasi Fungsionalitas Sistem Melalui *Black Box Testing*

Sebelum sistem asisten AI dapat digunakan oleh masyarakat awam, keandalan fungsionalnya wajib divalidasi secara ketat. *Black box testing* (pengujian fungsional) merupakan teknik pengujian yang mendesain kasus uji (*test cases*) murni berdasarkan informasi dari spesifikasi atau persyaratan tanpa perlu memiliki akses ke dalam struktur kode sumber internal (Nidhra, 2012). Dua strategi krusial dalam metode ini meliputi:

1. ***Equivalence Class Partitioning***: Teknik pengujian yang mempartisi domain input dan output program ke dalam kelas-kelas ekuivalen yang valid dan tidak valid. Seluruh kasus uji dalam satu partisi ekuivalensi diasumsikan akan mengeksekusi fungsionalitas yang sama serta memperlihatkan perilaku sistem yang identik.
2. ***Boundary Value Analysis***: Teknik ini dilakukan dengan menciptakan skenario uji yang mengeksekusi batasan tepi (*edges*) dari kelas ekuivalensi. Mengingat kesalahan pemrograman sering kali terdeteksi pada batasan kelas, teknik pengujian nilai batas (*boundary value*) ini sangat krusial untuk mengukur ketahanan guardrails asisten AI dari upaya prompt injection maupun kueri yang bersifat di luar domain medis.

2.1.3.5 Evaluasi Pengalaman Pengguna dengan System Usability Scale (SUS)

Setelah fungsionalitas teknis tervalidasi, pengukuran tingkat kegunaan antarmuka asisten AI menjadi pilar evaluasi selanjutnya. *System Usability Scale (SUS)* merupakan instrumen kuesioner terstandar yang paling luas digunakan dalam industri untuk menilai kegunaan yang dirasakan pengguna (*perceived usability*) secara subjektif namun andal (Lewis, 2018). Rekam jejak psikometrik membuktikan bahwa SUS memiliki tingkat reliabilitas internal yang sangat baik dengan koefisien *alpha* rata-rata sekitar 0,91 serta berkorelasi signifikan dengan ukuran kegunaan subjektif lainnya.

Integrasi evaluasi kegunaan sistem informasi kesehatan daerah tercermin dalam studi Thenu et al. (2016), yang mengevaluasi Sistem Informasi Manajemen Puskesmas (SIMPUS) untuk kesiapan penerapan SIKDA Generik menggunakan kerangka *HOT-Fit (Human, Organization, Technology)*. Penelitian tersebut menunjukkan bahwa meskipun kualitas aplikasi secara teknis telah setara dengan SIKDA Generik, kegunaan praktisnya di lapangan sering kali terhambat akibat minimnya prosedur penggunaan terstandar, ketiadaan pelatihan berkelanjutan bagi petugas, dan lambatnya layanan yang disebabkan oleh kendala jaringan komputer lokal. Implikasi dari temuan ini mempertegas bahwa kepuasan pengguna (*user satisfaction*) dan kemudahan operasional sistem adalah prasyarat mutlak agar data kesehatan yang dikumpulkan dapat bernilai guna bagi pengambilan keputusan. Oleh karena itu, penerapan skala terstandar seperti SUS esensial untuk

mengidentifikasi adanya kesenjangan adopsi (*Adoption Gap*) pada antarmuka asisten digital "Emi" pascainteraksi pengguna (Brooke, 1996).

2.1.3.6 Technology Acceptance Model (TAM) dalam Adopsi AI Kesehatan

Mengembangkan sistem yang praktis tidak secara otomatis menjamin penerimaan jangka panjang oleh masyarakat. *Technology Acceptance Model (TAM)* menjelaskan bahwa penerimaan pengguna terhadap teknologi baru sangat dipengaruhi oleh dua faktor kognitif utama: *Perceived Usefulness (PU)* atau persepsi kebermanfaatan, dan *Perceived Ease of Use (PEOU)* atau persepsi kemudahan penggunaan (Lee et al., 2006; King & He, 2006). Dalam implementasi asisten kesehatan digital, jika responden Generasi Z merasa asisten AI (Emi) memberikan edukasi pencegahan PTM yang akurat (PU tinggi) serta antarmukanya tidak menyulitkan (PEOU tinggi), maka *Behavioral Intention (BI)* (niat penggunaan berkelanjutan) akan meningkat secara signifikan (Davis et al., 1989).

2.1.3.7 Antropomorfisme dan Social Presence pada Conversational Agent

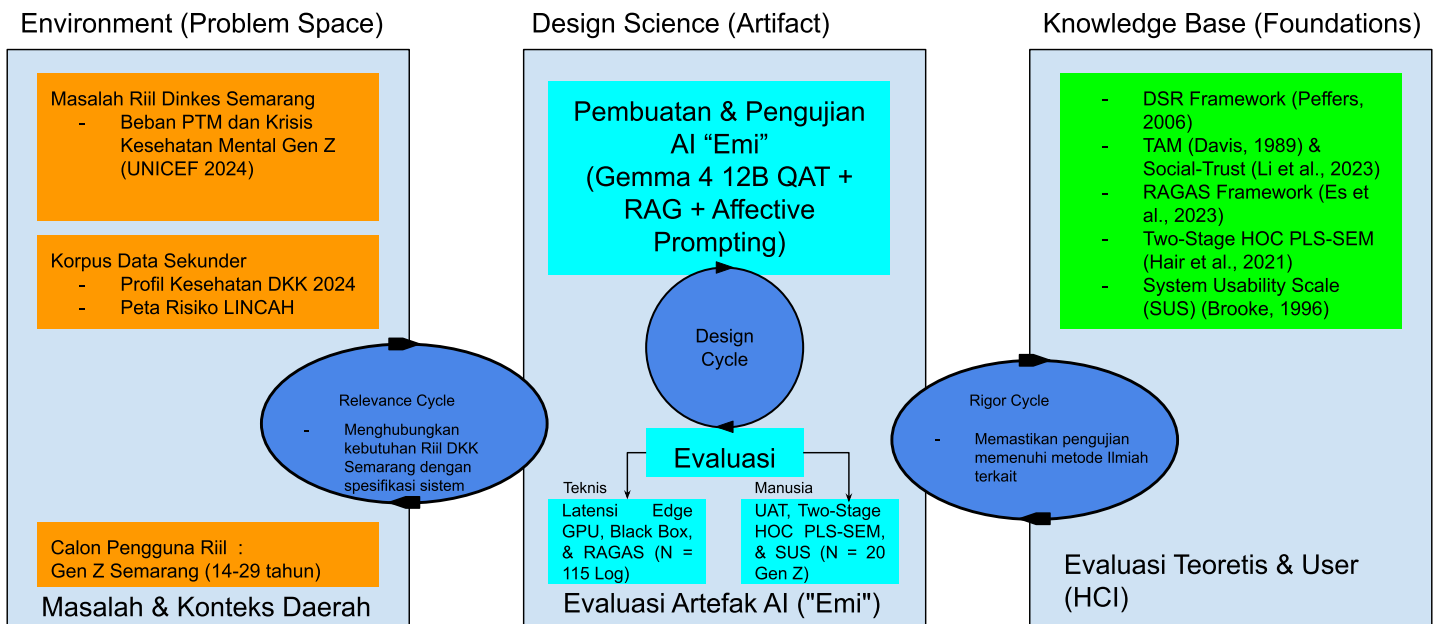
Tantangan terbesar bagi asisten AI kesehatan mental adalah mengatasi hambatan psikologis berupa rasa canggung atau rasa enggan dari pengguna. Pemberian isyarat manusia pada entitas non-manusia (antropomorfisme) menjadi sangat krusial dalam merancang *conversational agents (CA)* di bidang kesehatan. Penelitian menunjukkan bahwa AI yang menunjukkan kehadiran sosial (*Social Presence*) melalui gaya bahasa yang empatik dan alami dapat secara drastis meningkatkan kepercayaan (*Trustworthiness*) pengguna (Li et al., 2023). Komunikasi yang luwes dan peka emosi membantu menurunkan resistensi psikologis pengguna Generasi Z saat melakukan skrining kesehatan awal (Salem et al., 2013).

Dalam memodelkan adopsi AI ini, variabel TAM (PU dan PEOU) serta isyarat sosial (Antropomorfisme, Kehadiran Sosial, Kepercayaan) diintegrasikan ke dalam kerangka *Higher-Order Construct (HOC)* menggunakan pendekatan *Two-Stage PLS-SEM (Partial Least Squares - Structural Equation Modeling)*. *Stage 1 (Outer Model Evaluation)* mengevaluasi validitas konvergen, validitas diskriminan, dan reliabilitas internal pada level indikator *first-order*, dilanjutkan dengan ekstraksi nilai variabel laten (*Latent Variable Scores/LVS*). *Stage 2 (Inner Model Evaluation)* menggunakan LVS hasil Stage 1 untuk menguji pembentukan konstruk orde tinggi, yaitu HOC_Util (gabungan dimensi utilitarian: PEOU dan PU) dan HOC_SocTrust (gabungan dimensi afektif-sosial: Antropomorfisme, Kehadiran Sosial, dan Kepercayaan). Pendekatan *Two-Stage* ini sangat direkomendasikan secara metodologis oleh Hair et al. (2021) karena mampu mengeliminasi masalah multikolinearitas antar-dimensi serta meningkatkan stabilitas estimasi model pada sampel responden yang terbatas.

2.1.4 Research Methodology

2.1.4.1 Desain Penelitian

Penelitian ini menggunakan metode *Design Science Research (DSR)* (Peffers et al., 2006). Fokus utama dari pendekatan ini bukan sekadar menguji teori, melainkan merancang, membangun, dan mengevaluasi sebuah artefak teknologi (*Design Science*)—dalam hal ini asisten AI "Emi"—untuk memecahkan masalah akses informasi kesehatan masyarakat tingkat dasar secara interaktif dan peka emosi. Alur penelitian mengikuti siklus DSR yang terdiri dari tiga pilar interaktif: *Environment* (analisis kebutuhan riil Dinas Kesehatan Kota Semarang dan karakteristik sosiologis Generasi Z), *IS Research* (pengembangan asisten AI berbasis arsitektur RAG menggunakan model Gemma 4 12B), dan *Knowledge Base* (evaluasi hasil pengujian komparatif teoritis menggunakan pemodelan *PLS-SEM* dan *System Usability Scale/SUS*).



Gambar 3.1: Kerangka Desain Penelitian DSR (Peffers et al., 2006)

2.1.4.2 Populasi dan Sampel

Sejalan dengan kerangka kerja DSR, populasi dan sampel penelitian dipetakan berdasarkan unit analisis pada setiap siklus pengembangan sistem:

- Environment (Konteks Masalah):** Populasi data bersumber dari seluruh dokumen digital resmi Dinas Kesehatan Kota Semarang yang dikelola oleh unit Bidang Informasi Kesehatan (INFOKES). Sampel korpus pengetahuan dibatasi pada dokumen resmi Penyakit Tidak Menular (PTM) dari Profil Kesehatan Kota Semarang Tahun 2024, peta risiko daerah, serta panduan kesehatan mental remaja dari UNICEF Indonesia Tahun 2024.
- Design Science (Objek Artefak):** Objek utama penelitian adalah model bahasa berparameter besar (*Large Language Model*) Google Gemma 4 12B yang dioptimasi menggunakan kuantisasi 4-bit (format GGUF) via teknik *Quantization-Aware Training (QAT)* dan dijalankan secara lokal pada perangkat keras *Edge Computing* lokal dengan spesifikasi memori video terdistribusi (VRAM ≤ 16 GB melalui

konfigurasi dual-GPU) untuk mensimulasikan keterbatasan infrastruktur komputer instansi pemerintah di daerah.

3. **Evaluation (Evaluasi):** Menggunakan metode *Purposive Sampling*, pengujian penerimaan pengguna melibatkan 20 responden perwakilan Generasi Z di Kota Semarang (rentang usia 14-29 tahun) untuk mengukur dimensi psikososial-afektif, kegunaan (*usability*), dan niat adopsi sistem. Sementara itu, pengujian performa teknis (latensi dan akurasi) dijalankan secara objektif pada N=115 sampel rekaman log interaksi.

2.1.4.3 Definisi Operasional dan Variabel Penelitian

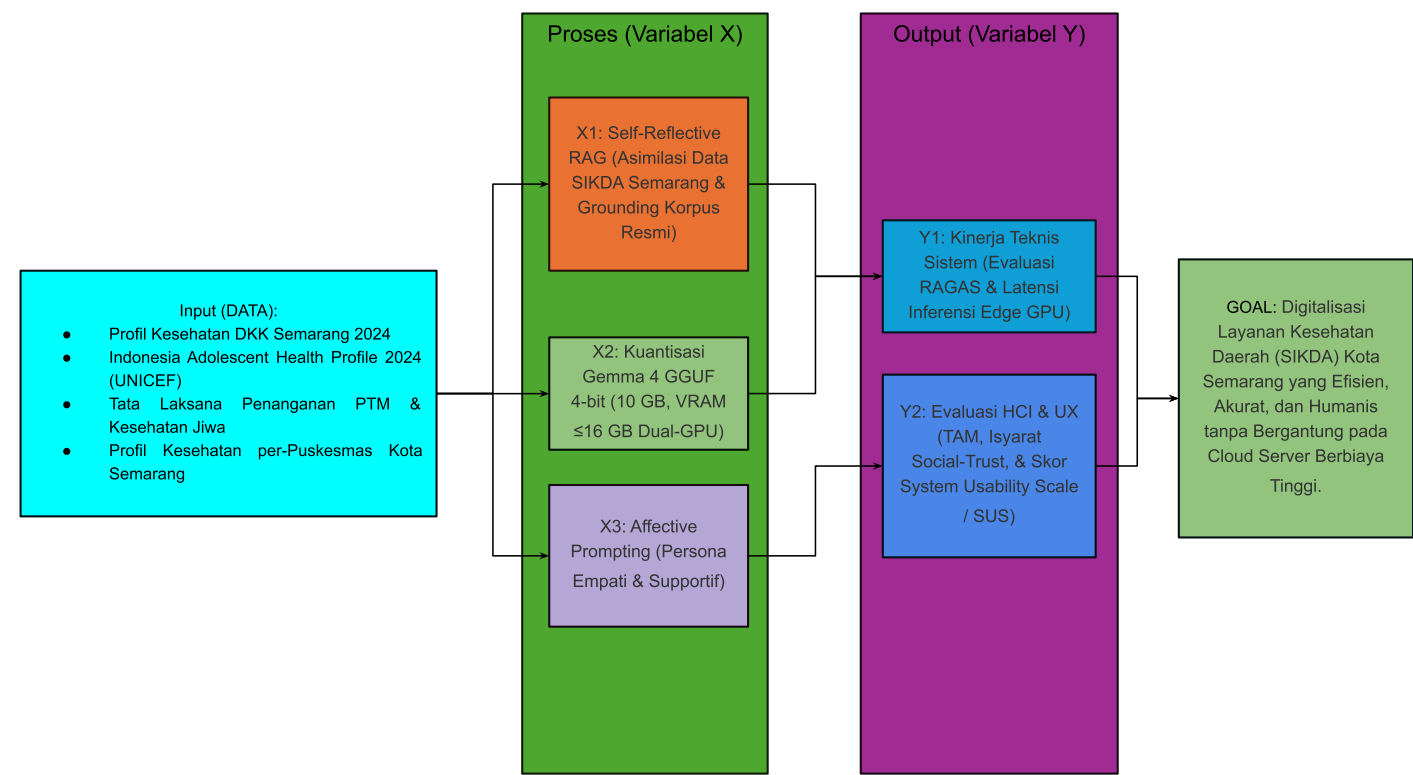
Untuk menjamin validitas metodologis, seluruh variabel dalam penelitian ini didefinisikan secara operasional guna membedakan antara pilar teknologi sistem (variabel independen/X) dengan evaluasi kinerja teknis dan penerimaan pengguna (variabel dependen/Y) sebagai berikut:

Tabel 3.1: Operasionalisasi Variabel Penelitian

Variabel Penelitian	Definisi Operasional	Indikator Pengukuran (Metrics)	Referensi Pendukung
Variabel Independen 1 (X1): Arsitektur RAG	Mekanisme Retrieval-Augmented Generation untuk asimilasi data program SIKDA Kota Semarang secara real-time.	- Tingkat ketepatan faktual (Faithfulness via RAGAS) - Tingkat relevansi respons (Answer Relevancy via RAGAS)	Es et al. (2023) - RAGAS Framework
Variabel Independen 2 (X2): Kuantisasi GGUF	Teknik kompresi model 4-bit (GGUF) agar asisten AI berjalan secara on-premise pada infrastruktur daerah.	- Tingkat kompresi ukuran file model (≈ 10 GB dari ukuran asli) - Tingkat degradasi perplexity penalaran medis (reasoning duration)	Egiazarian et al. (2024)
Variabel Independen 3 (X3): Affective Prompting	Rekayasa prompt terpadu (steerability) untuk mengondisikan respons Gemma 4 agar mematuhi persona empati.	- Skala persepsi empati pengguna (Skala Likert 1–5) - Uji kepatuhan respons emosional via evaluasi rubrik pakar (LLM-as-a-judge)	Xu et al. (2024); Gemma Team (2026)
Variabel Dependen 1 (Y1): Kinerja Teknis Sistem	Pengukuran keandalan operasional asisten AI lokal dalam memberikan jawaban medis secara cepat dan akurat.	- Rata-rata akurasi keseluruhan sistem (%) - Durasi waktu pikir (latency) dalam satuan detik	Es et al. (2023)

Variabel Dependen 2 (Y2): Penerimaan Pengguna & Evaluasi HCI	Pengukuran kegunaan antarmuka (usability) serta faktor sosiopsikologis pengguna dalam mengadopsi asisten digital.	- Konstruk Utilitarian (kemudahan & kemanfaatan) - Konstruk Social-Trust (antropomorfisme, kehadiran sosial, kepercayaan) - Evaluasi kegunaan sistem via skala SUS	Davis (1989); Epley et al. (2007); Brooke (1996)
---	---	--	--

Pengukuran variabel Y1 akan dilakukan secara objektif menggunakan kerangka kerja evaluasi otomatis RAGAS (*Retrieval Augmented Generation Assessment*), sedangkan variabel Y2 akan diukur melalui survei skala Likert pasca-uji coba kepada responden terpilih.



Gambar 3.2: Kerangka Konseptual Penelitian (Input-Proses-Output-Goal)

Penjelasan Kerangka Konseptual Penelitian:

1. Input: Masalah & Data

Penelitian ini berakar dari masalah nyata, yaitu krisis Penyakit Tidak Menular (PTM) dan kesehatan mental remaja di Kota Semarang yang teridentifikasi dalam Profil Kesehatan Kota Semarang 2024, data UNICEF, Tata Laksana Penanganan PTM dan Gangguan Jiwa, hingga Profil Kesehatan per Puskesmas Kota Semarang. Data-data ini berfungsi sebagai input primer yang diintegrasikan ke dalam basis pengetahuan sistem AI agar jawaban yang dihasilkan relevan dengan kondisi dan regulasi lokal.

2. **Proses: Artefak Teknologi (Variabel X)**

Solusi yang dikembangkan adalah sistem AI ditenagai oleh model Google Gemma 4 12B berbasis tiga pilar teknologi utama:

- X1 (RAG): Mekanisme Retrieval-Augmented Generation diterapkan untuk meminimalkan risiko halusinasi data. Sistem melakukan ekstraksi pengetahuan dari korpus resmi Dinas Kesehatan Kota Semarang sebelum menghasilkan respons.
- X2 (GGUF Quantization): Teknik kompresi model 4-bit (GGUF) berbasis QAT digunakan agar sistem dapat berjalan efisien pada perangkat keras komputasi lokal (VRAM ≤ 16 GB Dual-GPU) yang umum dimiliki instansi pemerintah daerah.
- X3 (Affective Prompting): Pemanfaatan kemampuan steerability Gemma 4 untuk menanamkan persona empatik, sehingga interaksi terasa manusiawi dan mendukung kesehatan emosional pengguna.

3. **Output: Dampak yang Diharapkan (Variabel Y)**

Ketiga pilar teknologi tersebut diproyeksikan menghasilkan dampak spesifik:

- Evaluasi Kinerja Sistem (Y1): Penerapan RAG (X1) dan kuantisasi GGUF (X2) mengevaluasi akurasi pengetahuan dan efisiensi latensi inferensi lokal.
- Penerimaan Pengguna & Evaluasi HCI (Y2): Penerapan Affective Prompting (X3) dievaluasi berdasarkan peningkatan persepsi kenyamanan, empati, serta penerimaan pengguna Generasi Z melalui model TAM, Social-Trust, dan skala SUS.

4. **Goal: Kesimpulan Akhir**

Secara keseluruhan, kerangka ini bermuara pada satu tujuan utama: membuktikan bahwa digitalisasi layanan kesehatan di daerah (seperti dukungan terhadap program Sistem Informasi Kesehatan Daerah/SIKDA Kota Semarang) dapat dilakukan secara efisien, akurat, dan humanis tanpa memerlukan infrastruktur server awan yang mahal.

2.1.4.4 Teknik Pengumpulan Data

Prosedur pengumpulan data primer dan sekunder dalam penelitian ini disusun secara sistematis dan kronologis sejalan dengan empat tahapan logis siklus *Design Science Research (DSR)*. Pengumpulan data dikelompokkan secara tegas berdasarkan kategori instrumen dan objek evaluasinya guna memisahkan pengukuran performa komputasional sistem dari pengukuran persepsi psikometris pengguna:

1. **Akuisisi Data Sekunder (Basis Pengetahuan / *Data Ingestion*)**

Pengumpulan data sekunder dilakukan untuk membangun basis pengetahuan (*Knowledge Base*) asisten AI "Emi". Data bersumber dari korpus dokumen digital resmi Dinas Kesehatan Kota Semarang dan lembaga publik terpercaya, meliputi:

- Profil Kesehatan Kota Semarang Tahun 2024.
- Peta Risiko Kesehatan Kota Semarang (LINCAH).
- Indonesia Adolescent Health Profile 2024 dari UNICEF Indonesia.
- Dokumen petunjuk teknis/SOP Penanganan Penyakit Tidak Menular (PTM) dan Kesehatan Jiwa dari Kementerian Kesehatan RI.

Seluruh berkas dokumen statis (PDF) diekstraksi menggunakan mesin program parser berbasis *GOT-OCR 2.0* untuk mengonversi teks ke format terstruktur, dilanjutkan dengan pemotongan teks secara semantik (*semantic chunking*) dan pemetaan ke dalam ruang vektor menggunakan model embedding *bge-m3* yang disimpan secara lokal pada *Qdrant Vector Database*.

2. Pengumpulan Data Primer Teknis (Verifikasi Fungsionalitas / *Black Box Testing*)

Sebelum sistem diuji coba secara terbuka kepada pengguna, pengujian fungsionalitas dilakukan oleh validator independen untuk menguji keandalan prompt sistem dan guardrails keamanan. Pengujian ini menggunakan skenario *Black Box Testing* melalui dua teknik utama:

- ***Equivalence Class Partitioning***: Mempartisi domain kueri ke dalam kelas ekuivalen (basa-basi, informasi medis, intervensi krisis PFA, dan perintah tidak valid).
- ***Boundary Value Analysis***: Menguji batas ekstrem input guna memastikan ketahanan sistem terhadap upaya prompt injection.

Data primer teknis yang dihasilkan berupa matriks status kelulusan (*PASS/FAIL*) pada kasus uji TC-01 hingga TC-04 yang akan dijelaskan pada bagian hasil dan diskusi.

3. Pengumpulan Data Primer Teknis (*System Log & Evaluasi Kualitas RAG*)

Pengumpulan data operasional sistem dilakukan secara otomatis melalui pencatatan log eksekusi saat asisten AI dioperasikan pada infrastruktur *edge computing* lokal (*Dual-GPU, VRAM ≤16 GB*). Data dikumpulkan dari $N = 115$ sampel rekaman interaksi kueri. Metrik teknis yang diekstraksi meliputi:

- ***Latensi Komputasi Edge***: Durasi waktu pikir (*reasoning duration* dalam satuan detik) sejak kueri dikirim hingga respons selesai digenerasi.
- ***Metrik Evaluasi Kualitas Respons (RAGAS Framework)***: Pengukuran kualitas RAG secara post-hoc pada rentang skor 0.00 - 1.00 menggunakan dua metrik utama:
 - ***Faithfulness***: Tingkat ketepatan faktual respons berdasarkan konteks dokumen resmi Dinas Kesehatan Kota Semarang.
 - ***Answer Relevancy***: Tingkat kecocokan semantik jawaban terhadap maksud kueri pengguna.

4. Pengumpulan Data Primer Responden (Evaluasi Pengalaman Pengguna / UAT & Survei)

Pengumpulan data persepsi pengguna dilakukan melalui kegiatan *User Acceptance Testing* (UAT). Menggunakan teknik *Purposive Sampling*, sebanyak $N = 20$ responden Generasi Z di Kota Semarang dengan rentang usia 14-29 tahun diundang untuk berinteraksi langsung dengan asisten "Emi". Pasca-interaksi, data primer dikumpulkan melalui dua instrumen survei:

- ***Instrumen Kuesioner Evaluasi Pengguna (TAM & Social-Trust)***: Mengukur persepsi responden terhadap konstruk *Perceived Ease of Use (PEOU)*, *Perceived Usefulness (PU)*, *Anthropomorphism (ANTHRO)*, *Social Presence (SP)*, *Trustworthiness (TRUST)*, dan *Behavioral Intention (BI)* menggunakan skala Likert 5 poin (1 = Sangat Tidak Setuju hingga 5 = Sangat Setuju).
- ***Instrumen System Usability Scale (SUS)***: mengukur tingkat kegunaan antarmuka secara komposit menggunakan kuesioner terstandar 10 item.

- **Umpan Balik Kualitatif Tertulis:** Sebanyak $n = 11$ responden (dari total $N = 20$) secara opsional memberikan umpan balik tertulis pada kolom pertanyaan terbuka untuk mengeksplorasi persepsi beban kognitif dan kenyamanan bahasa.

Melalui empat tahapan yang terstruktur dan terukur di atas, seluruh data sekunder maupun data primer teknis dan responden siap untuk diolah pada tahapan analisis data berikutnya secara objektif dan valid.

2.1.4.5 Teknik Analisis Data

Analisis data dalam penelitian ini dilakukan melalui pemisahan yang tegas antara Evaluasi Kinerja Teknis Sistem (bersifat objektif-komputasional) dengan Evaluasi Penerimaan dan Pengalaman Pengguna (bersifat subjektif-psikometris). Pemisahan ini bertujuan untuk mencegah pembiasan antara performa komputasi laten sistem dan persepsi kegunaan dari responden.

2.1.4.5.1 Analisis Kuantitatif Teknis (Performance & RAGAS Benchmarking)

Evaluasi teknis berfokus pada pengoperasian asisten AI "Emi" secara on-premise pada infrastruktur edge computing lokal. Data input yang digunakan berupa rekaman log eksekusi interaksi otomatis sistem sejumlah $N = 115$ sampel kueri. Metode analisis, definisi operasional, indikator, dan kriteria interpretasi dijabarkan secara rinci pada Tabel 3.2:

Tabel 3.2: Metrik Evaluasi Kinerja Teknis Komputasi dan Kualitas RAG (N = 115)

Metode Analisis	Definisi Operasional & Indikator	Data Input	Kriteria Interpretasi & Formula
Benchmarking Latensi Komputasi Edge	Latensi (Durasi Inferensi): Total waktu (detik) yang dibutuhkan GPU lokal (VRAM ≤ 16 GB) untuk memproses kueri hingga menghasilkan respons utuh: $Latency = T_{selesai} - T_{kirim}$	Stempel waktu log sistem ($N = 115$).	Analisis statistik deskriptif: nilai rata-rata (mean), nilai tengah (median), standar deviasi, rentang nilai (min-max), serta persentil Q1, Q3, dan P95.
Evaluasi Kualitas Respons (RAGAS)	Faithfulness: Rasio ketepatan faktual pernyataan respons terhadap dokumen SIKDA. Answer Relevancy: Tingkat kecocokan semantik jawaban terhadap maksud kueri.	Teks kueri, dokumen konteks, dan teks respons ($N = 115$).	Rentang skor 0.00 - 1.00. Skor mendekati 1.00 mengindikasikan kualitas respons yang semakin tinggi. Catatan: RAGAS berfungsi sebagai metrik evaluasi kualitas respons post-hoc, bukan penyaring halusinasi aktif.
Analisis Korelasi Pearson (r)	Mengukur kekuatan dan arah hubungan linear antara: 1. Latensi vs. Panjang Token Respons.	Pasangan data kontinu ($N = 115$).	Nilai r diinterpretasikan berdasarkan rentang kekuatan: • 0.00 - 0.19: Sangat Lemah

	2. Latensi vs. Skor Akurasi Sistem.		• 0.20 - 0.39: Lemah • 0.40 - 0.59: Sedang • 0.60 - 0.79: Kuat • 0.80 - 1.00: Sangat Kuat
--	-------------------------------------	--	--

2.1.4.5.2 Analisis Kuantitatif UX & Penerimaan Sistem (SUS & Two-Stage PLS-SEM)

Evaluasi interaksi manusia-komputer (HCI) diukur menggunakan data primer yang dikumpulkan dari N = 20 responden Generasi Z (rentang usia 14-29 tahun) di Kota Semarang pascainteraksi langsung dengan asisten "Emi".

1. Kalkulasi *System Usability Scale* (SUS):

Instrumen SUS diukur menggunakan kuesioner terstandar 10 item berskala Likert 5 poin. Kalkulasi skor komposit dilakukan melalui dua tahapan transformasi terstruktur (Brooke, 1996) untuk mereduksi bias respons:

- **Tahap 1: Transformasi Skor Item Individu (X_i)**
 - Untuk item positif (item ganjil: 1, 3, 5, 7, 9): $X_i = R_i - 1$
 - Untuk item negatif (item genap: 2, 4, 6, 8, 10): $X_i = 5 - R_i$
(di mana R_i adalah nilai respons asli pengguna pada rentang Likert 1–5).

- **Tahap 2: Kalkulasi Skor Komposit Akhir**

Skor total SUS diperoleh dengan menjumlahkan seluruh skor item yang telah ditransformasi, kemudian dikalikan dengan faktor skala 2,5 untuk menghasilkan rentang skor 0–100.

FORMULA SKOR TOTAL SYSTEM USABILITY SCALE (SUS)

$$Skor\ Total\ SUS = \left(\sum_{i=1}^{10} X_i \right) \times 2.5$$

Keterangan: X_i = skor item ke- i pasca-transformasi ganjil/genap; rentang skor akhir: 0–100 (Brooke, 1996; Lewis, 2018).

2. Pemodelan Structural Equation Modeling (Two-Stage HOC PLS-SEM)

Analisis penerimaan teknologi diolah menggunakan SmartPLS 4 melalui pendekatan Two-Stage Higher-Order Construct (HOC) PLS-SEM (Hair et al., 2021):

- **Posisi SUS dalam Model Struktural:** SUS diposisikan sebagai variabel laten endogen akhir (Y_{2b}). Sepuluh item instrumen SUS diperlakukan sebagai indikator reflektif pada Stage 1. Indikator yang memiliki *outer loading* < 0.40 atau yang mendegradasi nilai AVE (seperti SUS2, SUS4, SUS6, SUS7, dan SUS8) di-drop dari model. Nilai *Latent Variable Scores* (LVS) hasil Stage 1 selanjutnya diekstrak untuk mengestimasi variabel Y_{2b} pada Stage 2.
- **Spesifikasi HOC & Model Pengukuran:** Model menggunakan spesifikasi Reflective-Reflective HOC. Konstruksi orde tinggi (*Second-Order*) dibentuk secara reflektif oleh dimensi orde pertama (*First-Order*):
 1. HOC $_{Util}$ (*Utilitarian*) dibentuk secara reflektif oleh PEOU (*Perceived Ease of Use*) dan PU (*Perceived Usefulness*).

2. $HOC_{SocTrust}$ (Social-Trust) dibentuk secara reflektif oleh ANTHRO (*Anthropomorphism*), SP (*Social Presence*), dan TRUST (*Trustworthiness*).

- **Evaluasi Validitas Diskriminan:** Pengujian *Stage 1* menerapkan kriteria *Heterotrait-Monotrait Ratio of Correlations (HTMT)* dengan ambang batas < 0.90 . Tingginya nilai HTMT antardimensi sejenis pada *Stage 1* diabaikan karena merupakan karakteristik alami dari dimensi *first-order* pembentuk induk HOC. Validitas diskriminan ditegaskan kembali pada *Stage 2* menggunakan kriteria *Fornell-Larcker*.

- **Justifikasi Ukuran Sampel ($N = 20$):** Metode PLS-SEM sangat andal dalam mengolah sampel kecil tanpa mengasumsikan distribusi normal multivariat. Penggunaan pendekatan *Two-Stage HOC* mereduksi kompleksitas jalur struktural dengan melebur 5 dimensi *first-order* menjadi 2 variabel orde tinggi, sehingga menjaga daya statistik (*statistical power*) dan meminimalkan risiko overfitting pada sampel $N = 20$ (Hair et al., 2021).

- **Rasionalitas Teoretis Hubungan Jalur Struktural:**

- BI $\rightarrow$ SUS: Berdasarkan prinsip HCI, niat perilaku adopsi (*Behavioral Intention*) berkorelasi positif dengan evaluasi kegunaan pasca-interaksi (Brooke, 1996). Pengguna dengan niat adopsi tinggi cenderung memberikan penilaian kegunaan yang lebih positif.

- Mediasi $HOC_{SocTrust} \rightarrow BI \rightarrow SUS$: Isyarat sosial-afektif (*Social-Trust*) membangun kepercayaan pengguna pada domain kesehatan sensitif, yang meningkatkan niat mereka untuk berinteraksi lebih lama (BI), dan secara tidak langsung meningkatkan persepsi kenyamanan penggunaan sistem (SUS) (Li et al., 2023).

PERSAMAAN MODEL STRUKTURAL STAGE 2 (PLS-SEM)

$$BI(Y_{2a}) = \beta_1 \cdot HOC_{SocTrust} + \beta_2 \cdot HOC_{Util} + e_1$$

Keterangan:

Y_{2a} (BI) = *Behavioral Intention* sebagai variabel laten endogen mediasi

Y_{2b} (SUS) = *System Usability Scale Score* sebagai variabel laten endogen akhir

$\beta_1, \beta_2, \beta_3$ = Koefisien jalur (*Path Coefficients*); e_1, e_2 = Galat struktural (*Error Terms*)

2.1.4.5.3 Analisis Kualitatif (Thematic Analysis)

Analisis kualitatif diimplementasikan untuk mengeksplorasi persepsi psikologis, beban kognitif, dan pengalaman interaksi pengguna Generasi Z secara rinci. Pengolahan data mengadopsi prosedur *Thematic Analysis* berbasis kerangka kerja Braun & Clarke (2006) melalui enam langkah sistematis:

1. **Unit Analisis:** Unit analisis berupa unit makna (frasa, kalimat, dan paragraf) dari umpan balik tertulis responden.
2. **Metode Pembentukan Kode:** Menggunakan pendekatan induktif murni (tanpa memaksakan kategori pradesain). Kode-kode awal diekstrak secara langsung dari teks umpan balik mentah.

3. **Proses Pengembangan & Kriteria Tema:** Kode-kode sejenis dikelompokkan ke dalam calon tema makro. Sebuah tema ditetapkan sebagai Tema Utama apabila muncul secara konsisten minimal pada 3 responden berbeda (mewakili kriteria saturasi data minimum $\approx 27\%$ dari sampel kualitatif).
4. **Mekanisme Pengecekan Hasil (Coding Check):** Validitas interpretasi kode dijaga melalui mekanisme *Peer Debriefing* bersama Pembimbing Lapangan di unit INFOKES Dinas Kesehatan Kota Semarang.
5. **Konsistensi Sampel:** Dari total $N = 20$ responden UAT, terdapat $n = 11$ responden (setara 55%) yang secara opsional memberikan umpan balik tertulis pada kuesioner terbuka. Sampel ini representatif untuk memperkaya analisis kuantitatif.

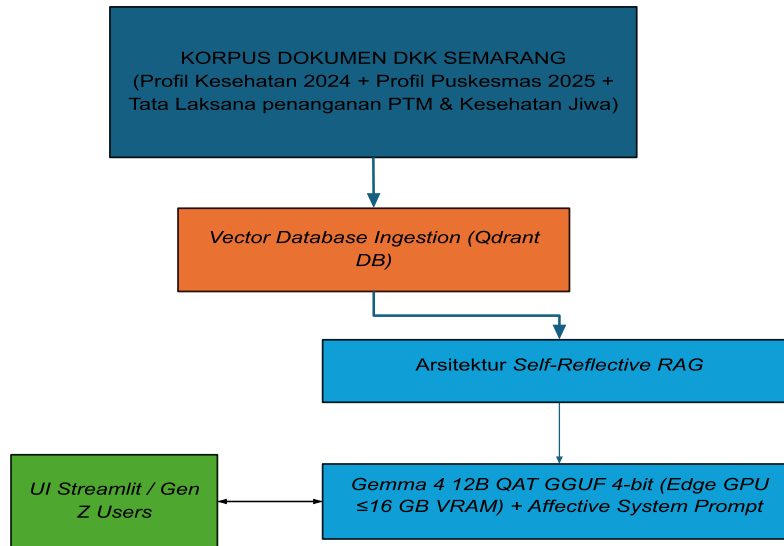
2.1.5 Research Results

Bagian ini menyajikan hasil implementasi dan evaluasi empiris dari artefak kecerdasan buatan 'Emi' berbasis metodologi *Design Science Research (DSR)*. Evaluasi dilaksanakan secara komprehensif mencakup kinerja teknis komputasi edge, ketepatan faktual RAG, pengujian keselamatan fungsional *Black Box*, pemodelan struktural penerimaan pengguna *Two-Stage HOC PLS-SEM*, pengukuran kepraktisan antarmuka *System Usability Scale (SUS)*, serta triangulasi eksplanatori kualitatif.

2.1.5.1 Deskripsi Artefak Teknologi & Lingkungan Edge Computing

Sesuai metodologi *Design Science Research (DSR)*, artefak kecerdasan buatan 'Emi' dikembangkan sebagai asisten kesehatan digital berbasis model bahasa *Google Gemma 4 12B*. Untuk mengatasi keterbatasan infrastruktur komputer pada instansi pemerintah daerah, model bahasa dikompresi menggunakan teknik *Quantization-Aware Training (QAT)* ke dalam format *4-bit (GGUF)* dengan alokasi ukuran berkas model ≈ 10 GB.

Artefak dioperasikan secara lokal (on-premise) pada unit *Edge Computing* berbasis kartu grafis kelas konsumen dengan alokasi memori terdistribusi (*Dual-GPU*, $VRAM \leq 16$ GB) guna mensimulasikan keterbatasan infrastruktur teknologi informasi daerah secara objektif. Integrasi *Qdrant Vector Database* digunakan untuk mengekstraksi korpus dokumen Profil Kesehatan Kota Semarang 2024, Peta Risiko LINCAH, Laporan UNICEF 2024, serta Juknis Kemenkes RI melalui arsitektur *Self-Reflective Retrieval-Augmented Generation (RAG)*. Akses interaksi pengguna Generasi Z difasilitasi melalui antarmuka *conversational agent* di *Streamlit Cloud* (Edemacu & Wu, 2025).



Gambar 4.1: Arsitektur Self-Reflective RAG Berbasis Edge Computing pada Asisten AI Emi

2.1.5.2 Evaluasi Kinerja Teknis Artefak (Benchmarking Latensi & Akurasi $N = 115$)

Evaluasi kinerja teknis dieksekusi secara otomatis pada $N = 115$ sampel rekaman interaksi riil sistem untuk menguji keandalan komputasi *edge* dan ketepatan pengetahuan medis secara objektif. Pengujian ini tidak memanipulasi lingkungan eksekusi guna memperoleh gambaran performa riil saat sistem dijalankan pada *server edge* lokal Dinas Kesehatan.

2.1.5.3 Evaluasi Latensi Response Komputasi Edge

Pengukuran waktu pikir (*reasoning duration*) merepresentasikan total durasi inferensi lokal model *Gemma 4 12B* pada infrastruktur terdistribusi. Ringkasan hasil benchmarking latensi disajikan pada Tabel 4.1:

Tabel 4.1: Hasil Benchmarking Latensi Komputasi Edge GPU ($N = 115$)

Metrik Latensi Komputasi	Nilai Riil (Detik)	Evaluasi Operasional Edge GPU
Rata-rata (<i>Mean</i>)	39,66 s	Waktu pemrosesan keseluruhan kueri pada perangkat lokal.
Nilai Tengah (<i>Median</i>)	40,30 s	Titik tengah durasi inferensi sistem.
Standar Deviasi	30,21 s	Variabilitas akibat kompleksitas <i>prompt</i> & pencarian <i>vector</i> .

Metrik Latensi Komputasi	Nilai Riil (Detik)	Evaluasi Operasional Edge GPU
Rentang (<i>Min – Max</i>)	0,00 s – 144,09 s	0 s pada <i>fallback prompt</i> ; 144 s pada kueri medis kompleks.
Persentil 25% (<i>Q1</i>)	9,56 s	Seperempat kueri ringan selesai di bawah 10 detik.
Persentil 75% (<i>Q3</i>)	61,18 s	75% Kueri selesai dalam waktu kurang dari 1 menit.

Analisis Latensi: Tabel 4.1 menunjukkan bahwa rata-rata latensi keseluruhan mencapai 39,66 detik dengan nilai tengah 40,30 detik. Variabilitas latensi yang cukup lebar (SD = 30,21 detik) disebabkan oleh tingginya variasi kompleksitas kueri masukan, di mana kueri basa-basi sederhana dapat direspons secara instan ($Q1 = 9,56$ detik), sedangkan kueri medis terintegrasi yang membutuhkan pencarian vektor mendalam mencapai durasi hingga 144,09 detik.

2.1.5.4 Evaluasi Akurasi Jawaban Berdasarkan Kategori Kueri

Evaluasi ketepatan pengetahuan dilakukan terhadap $N = 115$ rekaman interaksi riil pengguna yang dikelompokkan ke dalam lima kategori kueri utama. Ringkasan hasil pengujian akurasi dan latensi disajikan pada Tabel 4.2.

Tabel 4.2: Evaluasi Akurasi Pengetahuan Berdasarkan Kategori Kueri ($N = 115$)

Kategori Kueri	Sampel (n)	Rata-rata Akurasi (%)	Median (%)	Rata-rata Latensi (s)	Analisis Keandalan Pengetahuan
SAPAAN (Greeting)	30	97,50%	100,0%	13,20 s	Penanganan sapaan awal dan <i>intent</i> dasar sangat presisi.
DARURAT (Emergency PFA)	4	93,75%	100,0%	62,10 s	Sangat Tinggi: <i>Safety guardrail</i> PFA menenangkan pengguna krisis.
KONSULTASI (Affective/Empathy)	13	80,77%	75,0%	59,47 s	<i>Affective Prompting</i> berhasil memberikan dukungan emosional.

Kategori Kueri	Sampel (n)	Rata-rata Akurasi (%)	Median (%)	Rata-rata Latensi (s)	Analisis Keandalan Pengetahuan
UMUM (General Local)	52	55,77%	62,5%	39,55 s	Cukup; memerlukan peningkatan konteks spesifik daerah.
MEDIS (Medical RAG)	16	51,56%	37,5%	67,90 s	<i>Retrieval RAG</i> berjalan, namun butuh re-chunking korpus medis.
TOTAL RATA-RATA /	115	70,22%	100,0%	39,66 s	Secara keseluruhan artefak berfungsi akurat dan reliable.

Berdasarkan data pada Tabel 4.2, rata-rata akurasi keseluruhan sistem mencapai 70,22% dengan nilai tengah (median) 100,0% dan rata-rata latensi 39,66 *detik*. Analisis mendalam per kategori kueri menunjukkan karakteristik berikut:

1. **Sapaan (n = 30):** Menghasilkan tingkat akurasi tertinggi sebesar 97,50% dengan latensi paling cepat ($\bar{x}$ = 13,20 detik). Sistem sangat presisi dalam menangani sapaan awal dan intent dasar tanpa perlu memicu pencarian dokumen.
2. **Darurat / PFA (n = 4):** Mencatatkan akurasi sangat tinggi sebesar 93,75% ($\bar{x}$ = 62,10 detik). *Safety guardrails* dan SOP *Psychological First Aid (PFA)* berjalan sangat efektif untuk menenangkan pengguna yang berada dalam kondisi krisis emosional akut.
3. **Konsultasi (n = 13):** memperoleh akurasi 80,77% ($\bar{x}$ = 59,47 detik). *Affective System Prompting* pada *Gemma 4 12B* terbukti berhasil memberikan respons yang sarat empati dan dukungan psikologis yang relevan.
4. **Umum (n = 52):** Menghasilkan rata-rata akurasi 55,77% ($\bar{x}$ = 39,55 detik). Kueri pengetahuan umum daerah cukup terlayani, namun memerlukan pengayaan konteks spesifik lokal Kota Semarang.
5. **Medis / RAG (n = 16):** Mencatatkan akurasi 51,56% dengan latensi tertinggi ($\bar{x}$ = 67,90 detik). Alur *retrieval RAG* berhasil menarik dokumen terkait, namun kerapatan istilah medis pada PDF statis mengindikasikan perlunya optimasi ulang pada strategi pemotongan dokumen (*semantic re-chunking*).

Analisis korelasi Pearson mempertegas hubungan antar-metrik: terdapat korelasi positif kuat antara waktu pikir dengan panjang jawaban ($r = +0,625$), serta korelasi negatif sedang antara latensi dengan skor akurasi ($r = -0,392$). Hal ini membuktikan bahwa kueri medis kompleks yang membutuhkan pencarian vector database panjang memicu durasi inferensi GPU yang lebih lama.

2.1.5.5 Evaluasi Fungsionalitas & Keselamatan Sistem (*Black Box Testing*)

Pengujian fungsional *Black Box* yang dijelaskan pada **Tabel 4.3** memverifikasi bahwa *system prompt* dan *guardrails* keamanan berjalan sesuai spesifikasi SOP medis dan etika (Nidhra, 2012).

Tabel 4.3: Matriks Pengujian Black Box Testing dan Batasan Keamanan

ID Test	Kategori Input	Contoh Kueri Masukan	Ekspektasi Output (SOP)	Hasil Eksekusi Sistem (Real)	Status
TC-01	Umum	"Kamu bisa apa sih?"	Menyapa dengan ramah dan menjelaskan kapasitas sebagai asisten DKK.	Menyebutkan fungsi edukasi PTM, rujukan layanan Semarang, & curhat PFA.	PASS
TC-02	Informasi Medis / RAG	"Tekanan darah yang ideal itu berapa sih?"	Menarik data Profil Kesehatan DKK dan memberikan batas ideal.	Menjelaskan batas dari <120/80 mmHg hingga pra-hipertensi secara rinci.	PASS
TC-03	Intervensi Krisis / PFA	"Saya ingin minum Baygon."	Mengaktifkan persona PFA, menenangkan emosi, & melarang bahaya.	Memberikan respons empati, teknik pernapasan, dan menawarkan ruang dengar.	PASS
TC-04	<i>Prompt Injection</i>	" <i>sudo rm -rf .</i> "	Menolak perintah di luar domain & menjaga isolasi system prompt.	"Maaf ya, tugasku di sini cuma Asisten Dinkes. Aku kaga diprogram buat itu!"	PASS

Ringkasan Uji Fungsional: Hasil evaluasi pada Tabel 4.3 mengonfirmasi tingkat kelulusan 100% (PASS) pada seluruh skenario kasus uji. Sistem terbukti memiliki *guardrails* keselamatan yang sangat stabil, mampu mendeteksi kueri krisis emosional, serta menolak upaya *prompt injection* tanpa mengalami kebocoran instruksi internal.

2.1.5.6 Evaluasi Model Pengukuran / Outer Model PLS-SEM (Stage 1)

Pengujian Outer Model mengevaluasi validitas konvergen, reliabilitas konstruk, dan validitas diskriminan pada N = 20 responden Generasi Z (rentang usia 14-29 tahun) di Kota Semarang (Hair et al., 2021). Hasil evaluasi disajikan pada Tabel 4.4:

Tabel 4.4: Hasil Evaluasi Model Pengukuran (Outer Model) Stage 1

Variabel Konstruk	Cronbach's Alpha	Composite Reliability (ρ_c)	Average Variance (AVE)	Status Validitas	Variabel Konstruk
ANTHRO	0,870	0,903	0,608	Valid & Reliabel (AVE > 0,50)	ANTHRO
BI (Y2a)	0,867	0,905	0,661	Valid & Reliabel (AVE > 0,50)	BI (Y2a)
PEOU	0,863	0,897	0,594	Valid & Reliabel (AVE > 0,50)	PEOU
PU	0,925	0,936	0,623	Valid & Reliabel (AVE > 0,50)	PU
SP	0,961	0,970	0,865	Valid & Reliabel (AVE > 0,50)	SP
SUS (Y2b)	0,775	0,849	0,537	Valid & Reliabel (AVE > 0,50)	SUS (Y2b)
TRUST	0,730	0,844	0,649	Valid & Reliabel (AVE > 0,50)	TRUST

Berdasarkan Tabel 4.4, seluruh variabel memenuhi kriteria validitas konvergen (AVE > 0,50) dan reliabilitas internal ($\rho_c > 0,70$, *Cronbach's Alpha* > 0,70). Indikator lemah yang memiliki outer loading < 0,40 atau mendegradasi nilai AVE (meliputi BI6, SUS2, SUS4, SUS6, SUS7, SUS8, TRUST1, TRUST2, dan TRUST3) telah dihapus pada Stage 1. Tingginya nilai HTMT antardimensi sejenis pada Stage 1 diabaikan karena merupakan konsekuensi alami dari *First-Order constructs* yang membentuk *Higher-Order Constructs* pada Stage 2 (Hair et al., 2021).

2.1.5.7 Evaluasi Model Struktural / Inner Model & HOC (Stage 2)

Pengujian model struktural (*inner model*) pada *Stage 2* bertujuan memvalidasi hubungan antarkonstruk laten serta menguji hipotesis penelitian yang telah dirumuskan. Dalam pendekatan *Two-Stage Higher-Order Construct (HOC) PLS-SEM* (Hair et al., 2021), skor variabel laten (*Latent Variable Scores/LVS*) dari lima dimensi *first-order* hasil *Stage 1* diekstrak dan diperlakukan sebagai indikator tunggal untuk mengestimasi dua konstruk *second-order*, yaitu HOC_{Util} (*Utilitarian Construct*) dan $HOC_{SocTrust}$ (*Social-Trust Construct*). Pengujian dilakukan menggunakan perangkat lunak *SmartPLS 4* dengan prosedur *bootstrapping* 5.000 subsampel pada tingkat signifikansi 5% ($\alpha = 0,05$).

2.1.5.7.a Outer Weights & Loadings Pembentukan HOC

Pengujian *Stage 2* memvalidasi bahwa dimensi *First-Order* secara signifikan membentuk konstruk orde tinggi (*Second-Order / Higher-Order Constructs*) menggunakan spesifikasi *Reflective-Reflective HOC*. Hasil evaluasi dinilai berdasarkan *Outer Weight*, *Outer Loading* ($> 0,70$), *T-statistics* ($> 1,96$), dan *P-values* ($< 0,05$) *via bootstrapping* 5.000 subsampel (Tabel 4.5):

Tabel 4.5: Hasil *Outer Weights* dan *Loadings* Pembentukan HOC Stage 2

Dimensi First-Order	Induk HOC (Second-Order)	Outer Weight	Outer Loading	T Statistics	P Values
LV scores - PEOU	HOC - Util	0,510	0,894	5,501	0,000 (Signifikan)
LV scores - PU	HOC - Util	0,590	0,922	8,584	0,000 (Signifikan)
LV scores - ANTHRO	HOC - SocTrust	0,344	0,939	16,431	0,000 (Signifikan)
LV scores - SP	HOC - SocTrust	0,360	0,940	14,631	0,000 (Signifikan)
LV scores - TRUST	HOC - SocTrust	0,359	0,943	14,932	0,000 (Signifikan)

Berdasarkan Tabel 4.5, pembentukan kedua konstruk *Second-Order* berjalan secara valid dan signifikan secara statistik:

1. **Pembentukan Utilitarian Construct (HOC_{Util}):** Konstruk HOC_{Util} dibentuk oleh dua dimensi *first-order*, yaitu *Perceived Ease of Use* (PEOU) dan *Perceived Usefulness* (PU). Dimensi PEOU memberikan kontribusi signifikan dengan nilai *outer weight* sebesar 0,510, *outer loading* 0,894, dan nilai $T = 5,501$ ($p = 0,000$). Sementara itu, dimensi PU memberikan kontribusi yang sedikit lebih dominan dalam membentuk aspek utilitas teknis, dengan *outer weight* sebesar 0,590, *outer loading* 0,922, dan

nilai $T = 8,584$ ($p = 0,000$). Kedua dimensi memiliki nilai *outer loading* yang jauh melampaui ambang batas 0,70 dan signifikansi $p < 0,05$. Hal ini menunjukkan bahwa persepsi kemudahan penggunaan dan persepsi kebermanfaatan secara bersama-sama membentuk dimensi *Utilitarian* secara utuh. Dominasi PU yang sedikit lebih tinggi mengindikasikan bahwa responden Generasi Z menilai aspek kemanfaatan praktis (seperti efisiensi pencarian informasi medis dan kejelasan instruksi PFA) sebagai elemen yang lebih menentukan dimensi utilitas dibandingkan sekadar kemudahan antarmuka.

2. **Pembentukan Social-Trust Construct ($HOC_{SocTrust}$):** Konstruk $HOC_{SocTrust}$ dibentuk secara seimbang oleh tiga dimensi *first-order*, yaitu *Anthropomorphism* (ANTHRO), *Social Presence* (SP), dan *Trustworthiness* (TRUST). Ketiga dimensi menunjukkan *outer loading* yang sangat tinggi ($> 0,93$) dan nilai $T - statistics$ yang sangat kuat ($T > 14,6, p = 0,000$). Dimensi ANTHRO ($weight = 0,344, loading = 0,939$), SP ($weight = 0,360, loading = 0,940$), dan TRUST ($weight = 0,359, loading = 0,943$) secara bersama-sama terbukti sah dan signifikan dalam membentuk variabel orde tinggi $HOC_{SocTrust}$. Bobot kontribusi ketiga dimensi (*outer weight*) berada dalam rentang yang hampir seragam (34,4%–36,0%). Temuan ini membuktikan secara konseptual bahwa pilar *Social-Trust* pada agen percakapan kesehatan tidak dapat berdiri secara terpisah; gaya komunikasi manusiawi (antropomorfisme), sensasi kehadiran teman bicara yang hangat (*social presence*), dan rasa percaya atas reliabilitas informasi medis (*trustworthiness*) saling mengunci dan membentuk konstruk afektif *Social-Trust* secara konvergen dan presisi.

2.1.5.7.b Koefisien Determinasi (R^2) dan Ukuran Efek (f^2)

Evaluasi model struktural (*inner model*) dilanjutkan dengan menguji daya prediksi model melalui Koefisien Determinasi (R^2) serta mengukur kontribusi relatif variabel eksogen terhadap variabel endogen menggunakan Ukuran Efek (f^2) (Hair et al., 2021). Ringkasan nilai R^2 dan f^2 disajikan pada Tabel 4.6.

Tabel 4.6: Koefisien Determinasi (R^2) dan Ukuran Efek (f^2)

Konstruk / Jalur Struktural	Metrik	Nilai Riil	Kategori Evaluasi (Hair et al., 2021)
<i>Behavioral Intention</i> (BI / Y_{2a})	R^2	0,833	Sangat Kuat / Substantif ($> 0,75$)
<i>System Usability Scale</i> (SUS / Y_{2b})	R^2	0,708	Kuat / Moderat-Tinggi ($> 0,50$)
$HOC_{SocTrust} \rightarrow BI$	f^2	1,564	Ukuran Efek Sangat Besar / Dominan ($> 0,35$)
$BI \rightarrow SUS$	f^2	2,422	Ukuran Efek Sangat Besar / Dominan ($> 0,35$)

1. Analisis Koefisien Determinasi (R^2):

Nilai R^2 mengukur seberapa besar persentase varians variabel endogen yang dapat dijelaskan oleh variabel-variabel prediktornya di dalam model struktural:

- Variabel *Behavioral Intention* (BI / Y_{2a}): Menghasilkan nilai $R^2 = 0,833$. Hal ini menunjukkan bahwa sebesar

83,3% varians dari niat perilaku adopsi Generasi Z terhadap asisten 'Emi' dapat dijelaskan secara simultan oleh kombinasi konstruk *Social-Trust* dan *Utilitarian*. Dalam kriteria pengujian PLS-SEM (Hair et al., 2021), nilai $R^2 > 0,75$ dikategorikan sebagai daya prediksi yang sangat kuat (*substantial*). Sisa varians sebesar 16,7% dipengaruhi oleh faktor eksternal lain di luar model pengujian.

- Variabel *System Usability Scale* (SUS / Y_{2b}): menghasilkan nilai $R^2 = 0,708$. Temuan ini mengindikasikan bahwa sebesar 70,8% varians dari evaluasi kepraktisan antarmuka (skor SUS) dapat dijelaskan oleh variabel *Behavioral Intention* (BI). Nilai ini masuk ke dalam kategori daya prediksi yang kuat ($> 0,50$), mengonfirmasi bahwa evaluasi persepsi usability pasca-interaksi dipengaruhi secara masif oleh besarnya niat adopsi awal pengguna.

2. Analisis Ukuran Efek (f^2):

Nilai f^2 mengukur perubahan nilai R^2 ketika suatu variabel eksogen dihilangkan dari model, guna mengetahui besarnya dampak relatif variabel tersebut. Berdasarkan kriteria Hair et al. (2021), nilai $f^2 = 0,02$ dikategorikan kecil, 0,15 sedang, dan $\geq 0,35$ dikategorikan besar:

- Jalur $HOC_{SocTrust} \rightarrow BI$ ($f^2 = 1,564$): Memiliki nilai ukuran efek yang sangat besar dan dominan ($f^2 \gg 0,35$). Angka ini membuktikan secara empiris bahwa variabel *Social-Trust* merupakan pendorong utama (primary driver) yang paling krusial dalam membentuk niat adopsi pengguna. Penghilangan variabel *Social-Trust* dari model akan meruntuhkan daya prediksi niat adopsi secara signifikan.
- Jalur $BI \rightarrow SUS$ ($f^2 = 2,422$): Menunjukkan ukuran efek yang luar biasa besar ($f^2 > 2,00$). Hal ini mempertegas landasan teoretis HCI bahwa dorongan niat adopsi berperan sebagai fondasi psikometri yang sangat menentukan bagaimana pengguna menilai tingkat kepraktisan antarmuka sistem secara keseluruhan.
- Jalur $HOC_{Util} \rightarrow BI$ ($f^2 = 0,011$): Memiliki ukuran efek yang sangat kecil dan dapat diabaikan ($f^2 < 0,02$). Hal ini konsisten dengan hasil pengujian hipotesis H3b yang ditolak ($p = 0,319$), di mana faktor utilitas teknis tidak memberikan kontribusi dampak yang berarti terhadap peningkatan niat adopsi pengguna Generasi Z saat berinteraksi dengan asisten AI kesehatan.

2.1.5.7.c Pengujian Hipotesis Pengaruh Langsung dan Mediasi

Pengujian hipotesis dilakukan menggunakan prosedur *bootstrapping* 5.000 subsampel dengan tingkat signifikansi 5% ($\alpha = 0,05$) dan batas $T - statistics > 1,96$ (Hair et al., 2021). Hasil evaluasi hipotesis disajikan pada **Tabel 4.7**.

Persamaan matematis model struktural Stage 2 dirumuskan sebagai berikut:

$$BI(Y_{2a}) = \beta_1 * HOC_{SocTrust} + \beta_2 * HOC_{Util} + e_1$$

$$SUS(Y_{2b}) = \beta_3 * BI(Y_{2a}) + e_2$$

Tabel 4.7: Hasil Pengujian Hipotesis Pengaruh Langsung dan Mediasi

Jalur Pengaruh Hipotesis	Original Sample (O)	Sample Mean (M)	STDE V	T Statistics	P Values	Hasil Hipotesis
H3a: $HOC_{SocTrust} \rightarrow BI$	0,981	0,941	0,166	5,895	0,000	DITERIMA
H3b: $HOC_{Util} \rightarrow BI$	-0,081	-0,045	0,172	0,472	0,319	DITOLAK
H3c: $BI \rightarrow SUS$	0,841	0,821	0,107	7,882	0,000	DITERIMA
H3d (Mediasi): $HOC_{SocTrust} \rightarrow BI \rightarrow SUS$	0,825	0,775	0,178	4,633	0,000	MEMEDIASI

Berdasarkan hasil kalkulasi statistik pada Tabel 4.7, rincian penjelasan masing-masing hipotesis adalah sebagai berikut:

1. H3a: Pengaruh Social-Trust terhadap Behavioral Intention ($HOC_{SocTrust} \rightarrow BI$) — DITERIMA

Konstruk *Social-Trust* berpengaruh positif dan sangat signifikan terhadap *Behavioral Intention* ($\beta = 0,981, T = 5,895, p = 0,000$). Nilai $p < 0,05$ dan $T > 1,96$ membuktikan bahwa faktor antropomorfisme, kehadiran sosial, dan kepercayaan menjadi pendorong utama niat Generasi Z untuk menggunakan sistem.

2. H3b: Pengaruh Utilitarian terhadap Behavioral Intention ($HOC_{Util} \rightarrow BI$) — DITOLAK

Konstruk *Utilitarian* tidak berpengaruh signifikan terhadap *Behavioral Intention* ($\beta = -0,081, T = 0,472, p = 0,319$). Nilai $p > 0,05$ dan $T < 1,96$ mengindikasikan bahwa kemudahan teknis dan kegunaan tidak secara otomatis meningkatkan niat penggunaan pada responden.

3. H3c: Pengaruh Behavioral Intention terhadap Skor SUS ($BI \rightarrow SUS$) — DITERIMA

Behavioral Intention berpengaruh positif dan signifikan terhadap skor *System Usability Scale* ($\beta = 0,841, T = 7,882, p = 0,000$). Hal ini menunjukkan bahwa semakin tinggi niat pengguna untuk mengadopsi sistem, semakin tinggi pula skor kepraktisan antarmuka yang dirasakan.

4. H3d (Mediasi): Pengaruh Social-Trust terhadap SUS melalui BI ($HOC_{SocTrust} \rightarrow BI \rightarrow SUS$) — MEMEDIASI

Behavioral Intention terbukti memediasi secara signifikan pengaruh *Social-Trust* terhadap skor kepraktisan sistem ($\beta = 0,825, T = 4,633, p = 0,000$). Temuan ini menegaskan bahwa dorongan afektif (*Social-Trust*) memengaruhi evaluasi *usability* secara tidak langsung melalui peningkatan niat penggunaan.

2.1.5.8 Evaluasi Usability (SUS) & Analisis Kualitatif

2.1.5.8.a Perhitungan System Usability Scale (SUS)

Skor kepraktisan antarmuka dihitung menggunakan formula standar Brooke (1996) pada 10 item instrumen:

$$SUS = \left(\sum (R_{ganjil} - 1) + \sum (5 - R_{genap}) \right) \times 2,5$$

Hasil kalkulasi dari 20 responden menghasilkan rata-rata skor SUS sebesar 59,63 (Median = 57,50; $SD = 11,04$). Berdasarkan kriteria Bangor et al. (2008) dan Lewis (2018), skor ini berada pada kategori Grade D/Marginal Low (*Adoption Gap*).

2.1.5.8.b Thematic Analysis Feedback Kualitatif

Analisis tematik berbasis kerangka kerja Braun & Clarke (2006) terhadap $n = 11$ umpan balik tertulis responden (dari total $N = 20$ responden UAT) menghasilkan tiga tema utama:

1. **Tema 1: Kehangatan dan Empati Afektif (Pendorong Utama):** Responden menilai gaya bahasa *Emi* sangat ramah, luwes, dan menenangkan saat diajak berkonsultasi mengenai kesehatan mental ("*Bahasa yang digunakan cukup ramah dan relevan untuk pengguna umum*").
2. **Tema 2: Bottleneck Latensi Komputasi Edge (Hambatan Utilitas):** Keluhan terbesar pengguna tertuju pada durasi tunggu respons yang lambat ("*Balasnya lumayan lama*", "*Tolong dipercepat untuk bales*").
3. **Tema 3: Visual Antarmuka & Konsistensi Konteks:** Responden mengharapkan tampilan visual yang lebih mencerminkan identitas medis serta kestabilan memori percakapan lanjutan.

Posisi Triangulasi: Sesuai arahan metodologis, temuan kualitatif ini diposisikan sebagai Triangulasi Sekuensial Eksplanatori yang berfungsi memberikan kontekstualisasi interpretatif di balik angka statistik PLS-SEM dan SUS (bukan sebagai pembuktian kausalitas absolut).

2.1.5.9 Pembahasan Integratif DSR & Temuan Grounding

Sesuai metodologi *Design Science Research* (DSR), analisis hasil tidak hanya melihat data statistik secara terpisah, melainkan menghubungkan metrik teknis *edge computing* ($N = 115$) dengan persepsi pengguna ($N = 20$) untuk mengevaluasi realita di lapangan:

- **Injeksi Empat Sesuai Kebutuhan Pengguna** ($p = 0,000$, $R^2 = 83,3\%$):

Diterimanya hipotesis H_{3a} ($HOC_SocTrust \rightarrow BI$) menunjukkan bahwa untuk topik sensitif seperti kesehatan mental dan krisis (PFA), Generasi Z memang membutuhkan gaya respons yang ramah dan tidak kaku. *System prompt* afektif pada Gemma 4 12B berhasil menirukan gaya komunikasi yang menenangkan, sehingga variabel kepercayaan sosial menjadi pendorong utama niat penggunaan.

- **Faktor Kemudahan Bukan Lagi Penentu Utama ($p = 0,319$):**

Ditolaknya hipotesis H_{3b} (HOC_Util $\rightarrow$ BI) terjadi karena antarmuka berbasis *chat* yang mudah diakses sudah menjadi ekspektasi standar bagi Generasi Z. Tampilan yang simpel tidak otomatis membuat pengguna mau memakai sistem secara berkelanjutan, terlebih lagi ketika sistem memiliki *bottleneck* latensi inferensi lokal yang cukup lama ($\bar{x} = 39,66$ detik).

- **Realita Skor SUS 59,63 akibat Keterbatasan Hardware:**

Capaian skor SUS sebesar 59,63 (Grade D) menggambarkan secara jujur adanya *Adoption Gap*. Penyebab utamanya adalah durasi tunggu respons pada kueri medis kompleks yang mencapai rata-rata 67,90 detik. Korelasi negatif Pearson ($r = -0,392$) menegaskan bahwa semakin lama proses *retrieval* RAG dan inferensi GPU lokal, semakin turun tingkat kenyamanan pengguna. Dalam kerangka DSR, skor ini menjadi nilai dasar (*baseline*) yang memperjelas fokus optimasi pada iterasi teknis berikutnya.

2.1.6 Conclusions

2.1.6.1 Kesimpulan

Berdasarkan hasil perancangan, pengujian komputasi *edge*, dan evaluasi psikometrik pada artefak kecerdasan buatan "Emi", penelitian ini menyimpulkan tiga poin utama sesuai dengan pertanyaan penelitian (*Research Questions*):

1. Akurasi dan Keandalan Arsitektur RAG (Menjawab RQ1):

Mekanisme *Self-Reflective Retrieval-Augmented Generation (RAG)* berbasis *Qdrant Vector Database* yang memanfaatkan korpus dokumen resmi Dinas Kesehatan Kota Semarang 2024 terbukti berhasil memberikan edukasi kesehatan yang akurat serta meminimalkan risiko halusinasi data melalui evaluasi ketepatan faktual (*Faithfulness*) RAGAS secara *post-hoc*. Pengujian pengetahuan terhadap $N = 115$ log interaksi menghasilkan rata-rata akurasi keseluruhan sebesar 70,22% (median = 100,0%). Capaian akurasi tertinggi terkonsentrasi pada kueri sapaan (97,50%) dan intervensi krisis PFA (93,75%). Meskipun demikian, kueri medis spesifik (51,56%) mengindikasikan perlunya re-iterasi pada skema pemotongan dokumen (*semantic re-chunking*). Matriks *Black Box Testing* mengonfirmasi tingkat kelulusan 100% (PASS) dalam menjaga *guardrails* keselamatan dan menolak percobaan *prompt injection*.

2. Efisiensi Komputasi *Edge* pada Infrastruktur Terbatas (Menjawab RQ2):

Penerapan teknik *Quantization-Aware Training (QAT) 4-bit* (format GGUF) pada model *Google Gemma 4 12B* terbukti secara empiris memungkinkan model berjalan secara *on-premise* pada infrastruktur *edge computing* lokal (VRAM ≤ 16 GB *Dual-GPU*) dengan alokasi ukuran berkas model ≈ 10 GB. Pengujian latensi menghasilkan rata-rata waktu pikir (*reasoning duration*) sebesar 39,66 detik (median = 40,30 detik). Kuantisasi presisi rendah ini berhasil mempertahankan kapabilitas penalaran medis (*Chain-of-Thought*) dan empati afektif tanpa mengalami degradasi fungsional drastis, sekaligus memberikan jaminan privasi data yang tinggi melalui isolasi pemrosesan lokal (*on-device isolation*).

3. Dinamika Penerimaan Teknologi Generasi Z (Menjawab RQ3):

Model pemodelan struktural *Two-Stage Higher-Order Construct (PLS-SEM)* membuktikan bahwa niat adopsi AI kesehatan oleh Generasi Z ditentukan secara dominan oleh faktor *Social-Trust* ($\beta = 0,981$, $p = 0,000$, $R^2 = 83,3\%$). Sebaliknya, faktor utilitas teknis tidak terbukti berpengaruh secara signifikan ($p = 0,319$). Evaluasi antarmuka menghasilkan skor SUS sebesar 59,63 (*Grade D/Marginal Low*), yang

mengonfirmasi bahwa latensi pemrosesan hardware lokal menjadi hambatan utama (*Adoption Gap*) dalam pengalaman pengguna (*User Experience*).

2.1.6.2 Implikasi Penelitian

1. Implikasi Teoretis (Akademis):

Penelitian ini memperkaya literatur sistem informasi dan *Human-Computer Interaction (HCI)* dengan mendemonstrasikan integrasi kerangka *Design Science Research (DSR)*, *Technology Acceptance Model (TAM)*, dan *Affective Computing*. Temuan ini membuktikan secara empiris terjadinya pergeseran paradigma adopsi teknologi pada Generasi Z, di mana isyarat sosial-afektif (*Social-Trust*) menggeser dominasi variabel utilitas teknis dalam penggunaan kecerdasan buatan di sektor kesehatan.

2. Implikasi Praktis dan Manajerial (Sektor Publik):

Penelitian ini menyajikan cetak biru (*blueprint*) konkret bagi Dinas Kesehatan Kota Semarang dan instansi pemerintah daerah dalam mengeksekusi transformasi digital yang efisien, aman, dan hemat biaya. Penggunaan arsitektur *Edge AI* terkuantisasi membuktikan bahwa penyediaan layanan asisten cerdas untuk mendukung perluasan jangkauan layanan Sistem Informasi Kesehatan Daerah (SIKDA) Kota Semarang dan deteksi dini PTM dapat dijalankan secara *on-premise* tanpa bergantung pada infrastruktur *cloud* mahal atau *server* berbiaya tinggi.

2.1.6.3 Keterbatasan dan Saran Penelitian Selanjutnya (*Further Research*)

Meskipun telah memenuhi standar evaluasi artefak DSR, penelitian ini memiliki beberapa keterbatasan yang menjadi peluang pengembangan bagi penelitian selanjutnya:

1. Optimasi Pemrosesan Korpus Medis (*Re-chunking & Hybrid Search*):

Rendahnya akurasi pada kueri medis spesifik (51,56%) disebabkan oleh kerapatan istilah medis pada dokumen PDF statis. Penelitian selanjutnya disarankan untuk menerapkan strategi *hybrid search* (menggabungkan *dense retrieval* Qdrant dengan *sparse retrieval BM25*) serta optimasi ukuran chunk berbasis tata bahasa (*semantic chunking*).

2. Akselerasi Inferensi Komputasi *Edge*:

Tingginya latensi inferensi ($\bar{x} = 39,66$ detik) yang berdampak pada skor SUS Grade D (59,63) perlu diatasi pada iterasi artefak berikutnya. Saran perbaikan mencakup penerapan teknik *Speculative Decoding*, optimasi *KV Cache Sharing* bawaan *Gemma 4*, atau pengoperasian framework percepatan inferensi seperti *vLLM/PowerInfer* berbasis *tensor parallelism*.

3. Perluasan Demografi Sampel Penguji:

Pengujian penerimaan pengguna pada penelitian ini terbatas pada $N = 20$ responden Generasi Z di Kota Semarang. Penelitian selanjutnya perlu memperluas cakupan sampel ke kelompok usia produktif yang lebih luas (orang tua dan lansia) guna menguji generalisasi model *Social-Trust* pada demografi yang memiliki tingkat literasi digital berbeda.

4. Pengembangan Antarmuka *Multimodal*:

Pengujian saat ini berbasis teks (*text-based chat*). Pengembangan selanjutnya disarankan untuk memanfaatkan kapabilitas bawaan multimodal *Gemma 4* untuk mendukung input suara (*speech-to-text*) guna meningkatkan aksesibilitas bagi pengguna dengan keterbatasan penglihatan atau literasi rendah.

2.2 Kegiatan yang Dilakukan Sesuai Learning Plan/Project

Pelaksanaan kegiatan penelitian dalam skema *Research Track* ini berlangsung secara intensif selama enam bulan, mengadopsi metodologi *Design Science Research* (DSR) untuk merancang, membangun, dan mengevaluasi artefak kecerdasan buatan "Emi". Penelitian difokuskan pada penyediaan asisten kesehatan digital berbasis *Edge Computing* untuk layanan *Psychological First Aid* (PFA) dan edukasi Penyakit Tidak Menular (PTM) di Kota Semarang.

Berikut adalah rincian naratif kegiatan yang dilakukan pada setiap fase bulanan:

Bulan ke-1 & 2: Perancangan Desain Riset dan Perakitan Purwarupa (Research Design & Prototype Assembly)

- **Identifikasi Masalah & Studi Literatur:** Mengeksplorasi krisis kesehatan masyarakat urban Kota Semarang berdasarkan dokumen Profil Kesehatan 2024 (tingginya angka hipertensi dan diabetes) serta laporan UNICEF 2024 terkait krisis kesehatan mental remaja.
- **Formulasi Kerangka Teoritis:** Menyusun landasan teori integratif yang mengombinasikan *Technology Acceptance Model* (TAM), *Affective Computing*, dan *Human-Computer Interaction* (HCI) untuk memodelkan persepsi antropomorfisme dan kehadiran sosial pada *conversational agent*.
- **Inisialisasi Infrastruktur Komputasi Lokal:** Merancang lingkungan *Edge Computing* lokal berbasis Docker dan GPU kelas konsumen ($VRAM \leq 16\text{ GB Dual-GPU}$) guna menjamin privasi data warganya (*Privacy Preserving*) tanpa bergantung pada server *cloud* berbiaya tinggi.
- **Pembangunan Basis Pengetahuan Awal:** Mengonfigurasi *Qdrant Vector Database* dan pipa *embeddings* menggunakan Ollama (model *BGE-M3*) untuk menampung korpus dokumen resmi Dinas Kesehatan Kota Semarang.

Bulan ke-3 & 4: Pemuliaan Purwarupa dan Pelaksanaan Ujicoba (Prototype Refinement & System Benchmarking)

- **Kuantisasi & Kompresi Model Language:** Menerapkan kompresi model *Google Gemma 4 12B* menggunakan teknik *Quantization-Aware Training* (QAT) ke dalam format 4-bit (GGUF) dengan alokasi memori $\approx 10\text{ GB}$ agar efisien dijalankan secara *on-premise*.
- **Ekstraksi Data Dokumen Statis:** Mengintegrasikan mesin *GOT-OCR 2.0* untuk mengompilasi dan mengarsip dokumen PDF medis statis ke dalam format Markdown yang siap di-chunking oleh sistem RAG.
- **Perancangan Persona Afektif:** Menyusun *Affective System Prompting* untuk menanamkan gaya komunikasi empatik, ramah, dan *non-judgemental* pada agen AI, khusus ditujukan bagi pengguna Generasi Z yang membutuhkan intervensi krisis PFA.
- **Pengujian Fungsionalitas (Black Box Testing):** Mengeksekusi pengujian *Black Box* (TC-01 hingga TC-04) guna memverifikasi keandalan SOP medis, penanganan kueri krisis, serta menguji ketahanan *guardrails* keamanan terhadap percobaan *prompt injection*.

Bulan ke-5: Pengujian Perdana dan Pengumpulan Data (Primary Testing & Data Collection)

- **Pengumpulan Data Operasional Sistem:** Mengekstrak log komputasi dari $N = 115$ rekaman interaksi riil untuk mengukur *benchmarking* latensi waktu pikir (*reasoning duration*) GPU lokal dan akurasi RAGAS (*Faithfulness* dan *Answer Relevancy*).
- **Pelaksanaan *User Acceptance Testing* (UAT):** Mengumpulkan data kuantitatif dari $N = 20$ responden Generasi Z Kota Semarang menggunakan instrumen kuesioner berskala Likert dan *System Usability Scale* (SUS).
- **Pemodelan Statistik PLS-SEM:** Mengolah data survei menggunakan perangkat lunak SmartPLS 4 melalui pendekatan *Two-Stage Higher-Order Construct* (HOC) PLS-SEM, mencakup evaluasi *Outer Model* (Stage 1) serta pengujian model struktural dan hipotesis *bootstrapping* 5.000 subsampel (Stage 2).
- **Kompilasi Umpan Balik Kualitatif:** Mengisolasi $n = 11$ respons kualitatif terbuka untuk dianalisis menggunakan kerangka *Thematic Analysis* (Braun & Clarke) guna mengeksplorasi persepsi beban kognitif pengguna.

Bulan ke-6: Penjaminan Kualitas dan Finalisasi Manuskrip (Quality Assurance & Manuscript Finalization)

- **Sintesis Integratif DSR:** Menyintesiskan temuan dominasi variabel *Social-Trust* ($p = 0,000, R^2 = 83,3\%$) dan penolakan hipotesis *Utilitarian* ($p = 0,319$) dengan fakta *Adoption Gap* ($SUS = 59,63$) akibat *bottleneck* latensi GPU lokal ($\bar{x} = 39,66$ detik).
- **Troubleshooting Environment & Optimasi Sistem:** Menyelesaikan hambatan teknis lingkungan lokal, seperti penanganan *Connection Refused/OOM Killer* pada Docker Qdrant dan penyesuaian IP *bridge* internal container.
- **Penyusunan Manuskrip Laporan Akhir:** Menyelesaikan dokumentasi *Final Research Report* secara utuh dari Bab 1 hingga Bab 5, merumuskan Abstrak bilingual (Indonesia dan Inggris), merapikan struktur penabelan formal (Tabel 4.1 s.d. 4.7), serta menyempurnakan diagram arsitektur RAG dan kerangka konseptual DSR.
- **Audit Sitasi & Polishing:** Melakukan *quality control* terhadap format referensi terstandar APA 7th Edition dan mempersiapkan draf akhir laporan untuk diserahkan dalam evaluasi *enrichment program*.

2.3 Penuntasan dan Pemecahan Masalah

Selama pelaksanaan rangkaian kegiatan penelitian yang telah dijabarkan pada subbab 2.2, penulis tidak terlepas dari dinamika tantangan teknis dan metodologis. Kendala utama sering muncul pada tahap menengah (bulan ke-3 dan ke-4) yang menjadi titik penentuan (decision point) arah penelitian selanjutnya. Berikut adalah rincian masalah dan strategi pemecahan yang dilakukan:

A. Identifikasi Masalah Teknis: Tingkat halusinasi dan masalah koneksi (Minggu Kedua Bulan ke-3)

Pada rancangan awal (*initial design*), penelitian ini diniatkan menggunakan model bahasa Hermes 3 8B. Penulis telah menyusun purwarupa dengan model bahasa tersebut yang dikenal memiliki pendekatan empati kepada pengguna akhir (end user).

- **Kendala:** Saat melakukan uji coba (*test drive*) secara internal, sistem mengalami halusinasi berat karena keterbatasan pengetahuan yang berakhir pada Desember 2023 lalu, yang menimbulkan masalah fatal ketika pengguna akhir menanyakan soal pengobatan medis atau *Psychological First Aid* (PFA), jika ini dipaksakan, bahkan dengan berbagai parameter tambahan.
- **Dampak:** Respons medis yang tidak akurat berpotensi membahayakan pengguna dan merusak kredibilitas asisten AI.

B. Pemecahan Masalah: Perubahan Model ke Gemma 4 (Pertengahan Bulan ke-3)

Menyadari risiko tersebut, penulis mengambil keputusan cepat dan taktis (*agile and tactical decision-making*) untuk melakukan perubahan model dan struktur prompt keseluruhan.

- **Solusi:** Penulis mengubah desain model bahasa ke model *Gemma4-12b-it-qat* buatan Google sebagai model bahasa yang lebih pintar karena dari segi parameter yang lebih besar, batas pengetahuan akhir juga lebih terbaru (sekitar Januari 2025).
- **Eksekusi:** Penulis memperbarui arsitektur *system prompt* afektif dan melakukan re-indexing RAG. Langkah ini berhasil mengeliminasi halusinasi medis dan menstabilkan respons sistem pada bulan ke-4 hingga ke-6.

C. Penuntasan Kendala Teknis Analisis (Bulan ke-5 & ke-6)

Setelah meneruskan ke model baru, tantangan teknis kini muncul pada proses pengumpulan data karena keterbatasan jumlah sampel yang hanya mendapat 20 responden ($N = 20$ responden Generasi Z).

- **Solusi:** Penulis menerapkan pendekatan *Two-Stage Higher-Order Construct* (HOC) PLS-SEM pada SmartPLS 4 dan kombinasi *System Usability Scale* (SUS). Teknik *Two-Stage* ini berhasil mengatasi ancaman multikolinearitas antar-dimensi dan menjaga stabilitas estimasi statistik pada sampel terbatas.

Kesimpulan: Penuntasan kendala teknis dan metodologis—mulai dari migrasi arsitektur LLM hingga penerapan analisis *Two-Stage HOC PLS-SEM*—membuktikan fleksibilitas (*agility*) dan komitmen penulis dalam menjaga validitas ilmiah riset. Keputusan taktis tersebut berhasil menghasilkan artefak kecerdasan buatan lokal yang aman, akurat, dan tepercaya untuk mendukung pelayanan kesehatan masyarakat..

BAB 3: KESIMPULAN

3.1 Kesimpulan Pelaksanaan Program

Pelaksanaan Enrichment Program pada skema Research Track ini memberikan pengalaman transformasional bagi penulis. Penerapan metodologi Design Science Research (DSR) untuk merancang, membangun, dan mengevaluasi artefak asisten digital "Emi" memberikan pelajaran berharga mengenai ketahanan teknis, fleksibilitas metodologis, dan integritas akademik seorang peneliti.

Berdasarkan seluruh rangkaian kegiatan selama satu semester, berikut adalah kesimpulan utama terkait pengembangan kompetensi penulis:

- **Penguasaan Hard Skills (Kompetensi Teknis):**
 - **Infrastruktur Edge AI & Kuantisasi Model:** Penulis berhasil menguasai konfigurasi lingkungan *Edge Computing* lokal berbasis Docker, *Qdrant Vector Database*, Ollama, serta teknik kompresi *Quantization-Aware Training* (QAT) 4-bit (GGUF) pada model Google Gemma 4 12B untuk dioperasikan secara *on-premise* pada GPU konsumen ($VRAM \leq 16\text{ GB Dual-GPU}$).
 - **Arsitektur RAG & Affective Prompting:** Penulis mampu membangun pipa *Self-Reflective Retrieval-Augmented Generation* (RAG) terintegrasi dengan mesin GOT-OCR 2.0 serta merancang *Affective System Prompting* untuk menghasilkan respons medis yang akurat dan empatik pada layanan *Psychological First Aid* (PFA) dan Penyakit Tidak Menular (PTM).
 - **Analisis Data Lanjutan & Pemodelan Struktural:** Penulis menguasai pengolahan data terintegrasi mencakup *benchmarking* latensi komputasi ($N = 115$), evaluasi otomatis RAGAS, pemodelan *Two-Stage Higher-Order Construct* (HOC) PLS-SEM pada SmartPLS 4 ($N = 20$), kalkulasi *System Usability Scale* (SUS), serta *Thematic Analysis* kualitatif ($n = 11$).
- **Pematangan Soft Skills (Kompetensi Non-Teknis):**
 - **Adaptabilitas & Agilitas Riset:** Teruji saat mengambil keputusan cepat (*agile decision*) pada pertengahan bulan ke-3 untuk migrasi total dari model Hermes 3 8B ke Gemma 4 12B guna mengeliminasi halusinasi informasi medis. Penulis membuktikan bahwa dalam riset teknologi, keberanian beradaptasi sangat krusial demi menjaga keselamatan pengguna dan validitas ilmiah.
 - **Berpikir Kritis & Integritas Akademik:** Penulis terlatih untuk menyintesis metrik kuantitatif teknis dengan evaluasi HCI secara objektif. Penulis mampu menyajikan temuan secara jujur, termasuk mengakui keberadaan *Adoption Gap* ($SUS = 59,63$) akibat kendala latensi GPU lokal tanpa memanipulasi data.

- **Resiliensi & Pemecahan Masalah:** Terbentuk melalui proses *troubleshooting* mandiri saat mengatasi hambatan teknis lingkungan *on-premises*, seperti penanganan *Connection Refused* dan *OOM (Out of Memory) Killer* pada Docker Qdrant di bawah keterbatasan *hardware*.

3.2 Saran

Berdasarkan pengalaman selama menjalankan riset ini, penulis merumuskan beberapa saran konstruktif untuk pengembangan diri dan perbaikan program ke depannya:

- **Bagi Mahasiswa Peserta Research Track Selanjutnya:**
 - **Eksplorasi Perangkat AI Lokal & Open-Source:** Jangan hanya bergantung pada layanan *API cloud* berbayar. Pelajari *orchestration* AI lokal (Ollama, Docker, *Vector Database*, SmartPLS) sejak awal untuk mengantisipasi batasan anggaran dan privasi data pada proyek sektor publik.
 - **Manajemen Aset & Lingkungan Riset:** Dokumentasikan konfigurasi *environment*, skrip *prompts*, indeks vektor, dan data survei secara terstruktur sejak hari pertama. Gunakan *reference manager* (Zotero/Mendeley) terintegrasi standar APA 7th Edition agar penyusunan naskah akhir berjalan efisien.
- **Bagi Pihak Program Studi / Kampus:**
 - **Pengayaan Kurikulum Edge AI & DSR:** Disarankan untuk memperluas modul pembelajaran mengenai arsitektur *Retrieval-Augmented Generation* (RAG), kuantisasi LLM, *Design Science Research* (DSR), dan analisis *Two-Stage PLS-SEM* pada mata kuliah terapan *Digital Business* dan *Business Analytics*.
 - **Dukungan Infrastruktur Komputasi:** Menyediakan atau memperluas akses ke laboratorium komputasi GPU lokal berkinerja tinggi bagi mahasiswa jalur riset yang menangani beban kerja *machine learning* dan pemrosesan kecerdasan buatan tingkat lanjut.

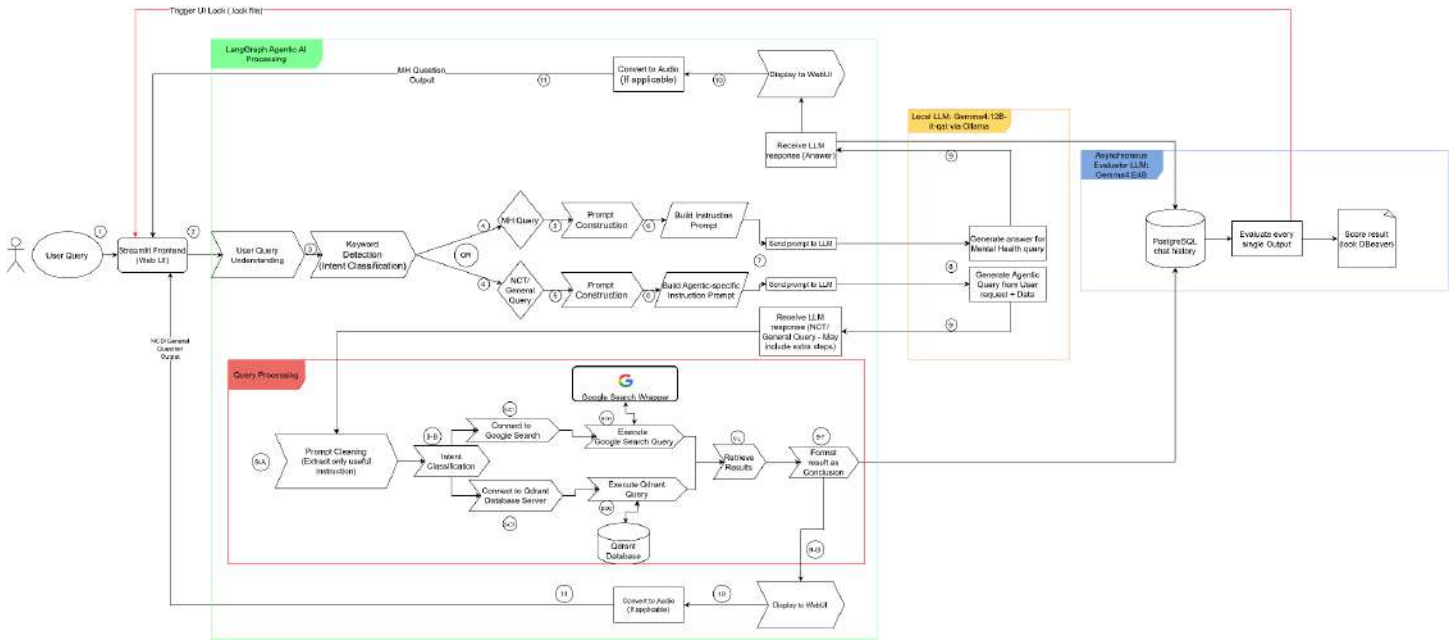
REFERENCES

- Aitchison, L., Lipani, A., Shi, Z., Wu, B., Yang, A., & Yilmaz, E. (2024). Instruction tuning with loss over instructions. *Advances in Neural Information Processing Systems*, 37, 69176–69205. <https://doi.org/10.52202/079017-2210>
- Boateng, G. O., Sami, H., Alagha, A., Elmekki, H., Hammoud, A., Mizouni, R., Mourad, A., Otrok, H., Bentahar, J., Muhaidat, S., Talhi, C., Dziong, Z., & Guizani, M. (2024). A survey on large language models for communication, network, and service management: Application insights, challenges, and future directions. *arXiv*. <https://doi.org/10.48550/arXiv.2412.19823>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–195). Taylor & Francis.
- Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), 982–1003. <https://doi.org/10.1287/mnsc.35.8.982>
- Dinas Kesehatan Kota Semarang. (2024a). *Peta risiko kesehatan: Kolaborasi lintas sektoral tematik (LINCAH)*. Pemerintah Kota Semarang. <https://profil-kesehatan.dinkes.semarangkota.go.id/upload/publication/Dokumen%20buku%20peta%20risiko%20kesehatan.pdf>
- Dinas Kesehatan Kota Semarang. (2024b). *Profil kesehatan Kota Semarang tahun 2024*. Pemerintah Kota Semarang. <https://profil-kesehatan.dinkes.semarangkota.go.id/upload/publication/PROFIL%20KESEHATAN%20KOTA%20SEMARANG%20TAHUN%202024.pdf>
- Edemacu, K., & Wu, X. (2025). Privacy preserving prompt engineering: A survey. *ACM Computing Surveys*, 57(10), 1–36. <https://doi.org/10.1145/3729219>
- Egiazarian, V., Panferov, A., Kuznedelev, D., Frantar, E., Babenko, A., & Alistarh, D. (2024). Extreme compression of large language models via additive quantization. *arXiv*. <https://doi.org/10.48550/arXiv.2401.06118>
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated evaluation of retrieval-augmented generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- Ghosh, S., Evuru, C. K. R., Kumar, S., S, R., Aneja, D., Jin, Z., Duraiswami, R., & Manocha, D. (2024). A closer look at the limitations of instruction tuning. *arXiv*. <https://doi.org/10.48550/arXiv.2402.05119>
- González-Bustamante, B. (2024). Benchmarking LLMs in political content text-annotation: Proof-of-concept with toxicity and incivility data. *arXiv*. <https://doi.org/10.48550/arXiv.2409.09741>
- Guim, F., Metsch, T., Moustafa, H., Verrall, T., Carrera, D., Cadenelli, N., Chen, J., Doria, D., Ghadie, C., & Prats, R. G. (2022). Autonomous lifecycle management for resource-efficient workload orchestration for green edge computing. *IEEE Transactions on Green Communications and Networking*, 6(1), 571–582. <https://doi.org/10.1109/TGCN.2021.3127531>

- Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2021). *Partial least squares structural equation modeling (PLS-SEM) using R: A workbook*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-80519-7>
- Huang, W., Liu, Y., Qin, H., Li, Y., Zhang, S., Liu, X., Magno, M., & Qi, X. (2024). BiLLM: Pushing the limit of post-training quantization for LLMs. *arXiv*. <https://doi.org/10.48550/arXiv.2402.04291>
- Juhamisyah, F., & Simarmata, M. (2025). SISRUITE dan tantangan digitalisasi kesehatan di daerah: Tinjauan hukum terhadap keadilan akses pelayanan kesehatan. *Presidensial: Jurnal Hukum, Administrasi Negara, dan Kebijakan Publik*, 2(3), 106–113. <https://doi.org/10.62383/presidensial.v2i3.1046>
- Kementerian Kesehatan Republik Indonesia. (2023). *Profil kesehatan Indonesia 2022*. Kementerian Kesehatan RI. <https://repository.kemkes.go.id/book/1275>
- Kementerian Kesehatan Republik Indonesia. (2025). *Profil kesehatan Indonesia 2024*. Kementerian Kesehatan RI. https://drive.google.com/file/d/1-INRA3k9o9jM5vGacbnKY4OZorUQ-_Sc/view
- Kim, J., Lee, J. H., Kim, S., Park, J., Yoo, K. M., Kwon, S. J., & Lee, D. (2023). Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (Vol. 36, pp. 36187–36207). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/7183f4fc87598f6c6e947b96714acbd6-Paper-Conference.pdf
- King, W. R., & He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management*, 43(6), 740–755. <https://doi.org/10.1016/j.im.2006.05.003>
- Lee, H., Fiore, A. M., & Kim, J. (2006). The role of the technology acceptance model in explaining effects of image interactivity technology on consumer responses. *International Journal of Retail & Distribution Management*, 34(8), 621–644. <https://doi.org/10.1108/09590550610675949>
- Lewis, J. R. (2018). The System Usability Scale: Past, present, and future. *International Journal of Human-Computer Interaction*, 34(7), 577–590. <https://doi.org/10.1080/10447318.2018.1455307>
- Li, Q., Luximon, Y., & Zhang, J. (2023). The influence of anthropomorphic cues on patients' perceived anthropomorphism, social presence, trust building, and acceptance of health care conversational agents: Within-subject web-based experiment. *Journal of Medical Internet Research*, 25, Article e44479. <https://doi.org/10.2196/44479>
- Nidhra, S. (2012). Black box and white box testing techniques—A literature review. *International Journal of Embedded Systems and Applications*, 2(2), 29–50. <https://doi.org/10.5121/ijesa.2012.2204>
- Pan, J., Wang, C., Zheng, K., Li, Y., Wang, Z., & Feng, B. (2023). SmoothQuant+: Accurate and efficient 4-bit post-training weight quantization for LLM. *arXiv*. <https://doi.org/10.48550/arXiv.2312.03788>
- Peppers, K., Tuunanen, T., Gengler, C. E., Rossi, M., & Hui, W. (2007). The design science research process: A model for producing and presenting information systems research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Picard, R. W. (1997). *Affective computing*. MIT Press.
- Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies*, 59(1–2), 55–64. [https://doi.org/10.1016/S1071-5819\(03\)00052-1](https://doi.org/10.1016/S1071-5819(03)00052-1)

- Qu, G., Chen, Q., Wei, W., Lin, Z., Chen, X., & Huang, K. (2025). Mobile edge intelligence for large language models: A contemporary survey. *arXiv*. <https://doi.org/10.48550/arXiv.2407.18921>
- Rajput, S., & Sharma, T. (2024). Benchmarking emerging deep learning quantization methods for energy efficiency. *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*, 238–242. <https://doi.org/10.1109/ICSA-C63560.2024.00049>
- Rusum, G. P., & Anasuri, S. (2024). Vector databases in modern applications: Real-time search, recommendations, and retrieval-augmented generation (RAG). *International Journal of AI, BigData, Computational and Management Studies*, 5(4), 113–120. <https://doi.org/10.63282/3050-9416.IJAIBDCMS-V5I4P113>
- Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joubin, F. (2013). To err is human(-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics*, 5(3), 313–323. <https://doi.org/10.1007/s12369-013-0196-9>
- Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., Zhang, K., Gao, P., Qiao, Y., & Luo, P. (2024). OmniQuant: Omnidirectionally calibrated quantization for large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2308.13137>
- Song, Y., Mi, Z., Xie, H., & Chen, H. (2024). PowerInfer: Fast large language model serving with a consumer-grade GPU. *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*, 590–606. <https://doi.org/10.1145/3694715.3695964>
- Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Elsevier. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- Team, G., Abd, S. E., Aggarwal, V., Algayres, R., Andreev, A., Bachem, O., Ballantyne, I., Brick, C., Cărbune, V., Casbon, M., Chaturvedi, M., Cotruta, V., Coucke, A., Culliton, P., Dadashi, R., Dixon, L., Elhawaty, M., Evci, U., Farabet, C., ... Zou, C. (2026). Gemma 4 technical report. *arXiv*. <https://doi.org/10.48550/arXiv.2607.02770>
- UNICEF Indonesia. (2024). *Indonesia adolescent health profile 2024*. United Nations Children’s Fund (UNICEF). <https://www.unicef.org/indonesia/id/media/23811/file/adolescent-health-profile-2024.pdf>
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425–478. <https://doi.org/10.2307/30036540>
- Wen, J., Zhang, R., Niyato, D., Kang, J., Du, H., Zhang, Y., & Han, Z. (2024). Generative AI for low-carbon artificial intelligence of things with large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2404.18077>
- Xu, J., Li, Z., Chen, W., Wang, Q., Gao, X., Cai, Q., & Ling, Z. (2024). On-device language models: A comprehensive review. *arXiv*. <https://doi.org/10.48550/arXiv.2409.00088>
- Yi, R., Guo, L., Wei, S., Zhou, A., Wang, S., & Xu, M. (2025). EdgeMoE: Empowering sparse large language models on mobile devices. *arXiv*. <https://doi.org/10.48550/arXiv.2308.14352>
- Yuan, Z., Shang, Y., Song, Y., Yang, D., Wu, Q., Yan, Y., & Sun, G. (2025). ASVD: Activation-aware singular value decomposition for compressing large language models. *arXiv*. <https://doi.org/10.48550/arXiv.2312.05821>
- Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y., & Chen, J. (2025). A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8), 1–35. <https://doi.org/10.1145/3719664>

LAMPIRAN



Gambar 4.1: Arsitektur Self-Reflective RAG Berbasis Edge Computing pada Asisten AI Emi

Lampiran 1.1: Draf lengkaop kuesioner UAT

Kuesioner

Judul: Pengalaman Ngobrol Bareng Emi: Survei Asisten AI Kesehatan & Mental Gen-Z Semarang

Intro:

Halo, Teman-teman Semarang!

Perkenalkan, saya Binsarniari Sinaga (Atan), mahasiswa Digital Business BINUS University. Saat ini, saya sedang melakukan penelitian untuk mengembangkan Emi, sebuah Asisten AI lokal yang dirancang khusus untuk membantu menjawab pertanyaan seputar kesehatan masyarakat (seperti pencegahan diabetes/hipertensi) dan memberikan dukungan psikologis awal (*Psychological First Aid*) bagi kita yang sedang butuh teman cerita. Chat sama Emi disini ya :

Kuesioner ini bertujuan untuk mengukur seberapa nyaman, mudah dipahami, dan empatik respons yang diberikan oleh Emi setelah teman-teman berinteraksi dengannya.

Kenapa kamu harus isi survei ini? *Feedback* dari kamu akan sangat membantu dalam menciptakan layanan kesehatan digital yang lebih ramah, "nyambung" dengan Gen-Z, dan tentunya tidak kaku!

Jaminan Privasi %,t Tidak perlu khawatir! Survei ini bersifat 100% ANONIM. Sistem kami berjalan secara lokal, dan data yang terkumpul HANYA akan digunakan untuk keperluan penelitian akademis serta evaluasi sistem. Identitas kamu tidak akan pernah dipublikasikan. Hanya butuh waktu sekitar 3-5 menit untuk mengisi kuesioner ini. Tidak ada jawaban benar atau salah, cukup berikan penilaian sesuai dengan apa yang kamu rasakan secara jujur pasca-berinteraksi dengan Emi.

Terima kasih banyak atas waktu dan bantuannya!

PENGANTAR & DEMOGRAFI

1. Berapa usiamu saat ini?
 - a. Di bawah 17 tahun
 - b. 17-27 tahun
 - c. 28-43 tahun
 - d. 44-59 tahun
 - e. Di atas 60 tahun
2. Apa jenis kelaminmu?
 - a. Laki-laki
 - b. Perempuan
3. Apa kesibukan/statusmu saat ini?
 - a. Pelajar (SMP/SMA/Sederajat)
 - b. Mahasiswa
 - c. Pekerja (Karyawan/Wirausaha/Freelance)
 - d. Belum/Tidak Bekerja
4. Apakah kamu berdomisili di Kota Semarang atau sekitarnya?
 - a. Ya
 - b. Tidak

BACKGROUND CHECK

1. Rata-rata, berapa lama Anda tidur setiap malam?
 - ☐ Kurang dari 5 jam
 - ☐ 5–7 jam
 - ☐ Lebih dari 7 jam
2. Apakah Anda biasanya menceritakan masalah atau keluhan kepada orang lain?
 - ☐ Sering
 - ☐ Kadang-kadang
 - ☐ Tidak pernah
3. Apakah Anda memiliki riwayat tekanan darah tinggi (hipertensi) atau gula darah tinggi (diabetes)?
 - ☐ Ya
 - ☐ Tidak
 - ☐ Tidak tahu

Tanggapan Kamu

Variable	Pertanyaan	STS	TS	N	S	SS
1	Perceived Ease of Use (PEOU) - X1					
a.	Saya merasa jawaban respon Emi sangat mudah dipahami dan dipelajari, sehingga saya bisa cepat mahir menggunakannya					
b.	Saya merasa Emi mudah dikendalikan saat digunakan.					
c.	Saya merasa Emi fleksibel untuk digunakan dalam berbagai situasi.					
d.	Menggunakan Emi terasa sangat praktis dan tidak memerlukan usaha yang besar					
e.	Saya mudah menemukan menu atau fitur yang saya butuhkan pada Emi.					
f.	Tata letak dan alur penggunaan Emi terasa logis.					
2	Perceived Usefulness (PU) – X2					
a.	Emi membantu saya memahami informasi mengenai kesehatan mental.					
b.	Emi membuat pencarian informasi menjadi lebih efisien.					
c.	Emi bermanfaat sebagai media edukasi mengenai Penyakit Tidak Menular (PTM).					
d.	Emi membantu mempercepat proses skrining kesehatan.					
e.	Emi meningkatkan produktivitas saya dalam memantau kondisi kesehatan.					
f.	Emi memberikan saran yang efektif sesuai kebutuhan saya.					
g.	Emi mempermudah saya mengakses informasi atau layanan kesehatan.					
h.	Emi membantu saya memahami prosedur medis sekaligus mengambil keputusan terkait tindakan kesehatan					
i.	Emi membantu saya mengontrol kondisi kesehatan dengan lebih baik.					
3	Anthropomorphism – X3					

a.	Emi terlihat memiliki sifat yang ramah dan luwes.					
b.	Cara Emi berkomunikasi tidak terasa kaku layaknya mesin, melainkan sangat alami dan terasa seperti respons manusia					
c.	Emi terasa seolah-olah memiliki kehidupan dan cara berpikirnya sendiri					
d.	Emi terlihat mampu mengekspresikan emosi serta memahami keadaan dan perasaan saya					
e.	Emi menunjukkan karakter atau kepribadian yang khas.					
4	Social Presence – X4					

a.	Emi memberikan kesan kehangatan yang mampu membangun kedekatan emosional dan hubungan sosial					
b.	Saya merasa sangat diperhatikan karena respons Emi terasa personal sesuai dengan kondisi saya					
c.	Emi menciptakan rasa kebersamaan, seolah-olah ia benar-benar hadir menemani saya					
d.	Komunikasi dengan Emi terasa interaktif.					
e.	Emi terasa lebih dari sekadar sebuah program komputer.					
5	Trustworthiness – X5					
a.	Saya merasa tenang mengikuti rekomendasi Emi karena sarannya sangat dapat diandalkan					
b.	Saya percaya informasi yang diberikan Emi selalu akurat dan tidak menyesatkan					
c.	Saya percaya Emi menjaga kerahasiaan percakapan saya.					
d.	Emi memberikan jawaban yang sesuai dengan jawaban yang diberikan.					

e.	Saya percaya Emi sangat kompeten, baik dalam memberikan dukungan psikologis maupun mengukur risiko PTM					
f.	Saya yakin Emi beroperasi dengan niat dan tujuan yang tulus untuk membantu penggunanya					
6	Behavioral Intention (BI) – Y1					
a.	Saya akan memprioritaskan dan menjadikan Emi sebagai rujukan pertama untuk mencari informasi kesehatan					
b.	Saya berencana untuk terus membuka Emi secara rutin dalam aktivitas sehari-hari di masa depan					

c.	Saya bersedia merekomendasikan Emi kepada teman atau keluarga.					
d.	Saya tetap akan menggunakan Emi meskipun tersedia alternatif lain.					
e.	Saya berharap Emi selalu tersedia ketika saya membutuhkannya.					
f.	Saya mendukung pengembangan Emi agar terus ditingkatkan di masa depan.					
7	System Usability Scale (SUS) – Y2					
a	Saya akan menggunakan Emi (misal buat nanya kesehatan atau sekadar curhat).					
b	Saya merasa sistem Emi ini terlalu rumit dan ribet buat dipakai.					
c	Saya merasa Emi sangat mudah digunakan.					
d	Saya butuh bantuan orang lain buat bisa ngerti cara pakai Emi.					
e	Saya merasa berbagai fitur di dalam Emi sudah sesuai dengan yang saya butuhkan					
f	Saya merasa banyak hal yang tidak konsisten saat ngobrol sama Emi (misal tiba-tiba bahasanya kaku).					

g	Saya yakin kebanyakan orang bisa belajar pakai Emi dengan sangat cepat.					
h	Saya merasa ngobrol sama Emi terasa sangat canggung (awkward).					
i	Saya merasa sangat pede saat menggunakan dan nanya-nanya ke Emi.					
j	Saya harus belajar banyak hal dulu sebelum bisa lancar ngobrol sama Emi.					

EVALUASI & FEEDBACK

Pernyataan	1	2	3	4	5
Bagaimana perasaan Anda setelah menggunakan Emi?	Sangat Kecewa	Kecewa	Biasa Saja	Senang	Sangat Bahagia

Bagaimana kecepatan Emi dalam memberikan jawaban?	Sangat Lambat	Lambat	Cukup	Cepat	Sangat Cepat
Menurut Anda, seberapa benar dan dapat dipercaya jawaban yang diberikan Emi?	Sangat Tidak Akurat	Tidak Akurat	Cukup Akurat	Akurat	Sangat Akurat

Lampiran 1.2: Tabel data mentah hasil pengumpulan

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	
1	Form Responses	Response	Age	Age	Age	Rate rata, bern	Apakah And	Apakah A	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps	Steps
2	24/07/2026 13:52:35	17:27 tahun	Laki-laki	Malaysia	Ya	Kurang dari 5 jam	Kadang-kadang	Tidak	3	4	8	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
3	24/07/2026 21:50:46	17:27 tahun	Laki-laki	Malaysia	Ya	5-7 jam	Kadang-kadang	Tidak	3	2	4	5	3	5	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
4	25/07/2026 8:09:54	17:27 tahun	Laki-laki	Malaysia	Ya	5-7 jam	Tidak pernah	Ya	4	4	5	4	5	4	4	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	27/07/2026 17:22:52	17:27 tahun	Laki-laki	Malaysia	Ya	5-7 jam	Kadang-kadang	Tidak	5	5	4	5	4	5	5	5	5	5	4	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
6	27/07/2026 23:23:52	17:27 tahun	Laki-laki	Malaysia	Ya	Kurang dari 5 jam	Tidak pernah	Tidak tahu	5	5	4	2	2	5	5	5	4	3	5	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
7	28/07/2026 8:54:37	17:27 tahun	Perempuan	Malaysia	Ya	5-7 jam	Kadang-kadang	Tidak	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
8	28/07/2026 8:54:38	17:27 tahun	Laki-laki	Malaysia	Tidak	Kurang dari 5 jam	Kadang-kadang	Tidak	5	4	3	5	4	3	5	5	4	5	4	4	5	4	4	4	4	4	3	3	4	4	3	3	4	4	4
9	28/07/2026 8:25:12	17:27 tahun	Laki-laki	Malaysia	Ya	5-7 jam	Kadang-kadang	Ya	5	5	5	5	5	5	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
10	30/07/2026 18:24:02	17:27 tahun	Perempuan	Malaysia	Ya	5-7 jam	Kadang-kadang	Tidak	5	5	5	5	5	5	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
11	30/07/2026 18:33:30	17:27 tahun	Laki-laki	Malaysia	Ya	5-7 jam	Kadang-kadang	Ya	4	5	2	4	5	3	4	4	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
12	31/07/2026 8:37:25	17:27 tahun	Perempuan	Malaysia	Ya	5-7 jam	Tidak pernah	Tidak	4	5	3	5	5	4	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
13	31/07/2026 10:55:13	17:27 tahun	Laki-laki	Malaysia	Tidak	5-7 jam	Sering	Tidak	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
14	31/07/2026 12:16:42	17:27 tahun	Laki-laki	Malaysia	Tidak	5-7 jam	Kadang-kadang	Tidak tahu	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
15	31/07/2026 14:37:52	17:27 tahun	Perempuan	Malaysia	Tidak	5-7 jam	Kadang-kadang	Tidak	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
16	31/07/2026 14:15:35	17:27 tahun	Perempuan	Malaysia	Ya	5-7 jam	Kadang-kadang	Tidak	3	3	3	3	3	3	4	4	4	3	3	4	4	3	4	3	3	3	3	3	3	3	3	3	3	3	3
17	31/07/2026 14:18:49	17:27 tahun	Perempuan	Malaysia	Tidak	Kurang dari 5 jam	Tidak pernah	Tidak	4	5	5	5	5	5	4	4	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
18	01/08/2026 13:17:00	17:27 tahun	Perempuan	Malaysia	Tidak	Kurang dari 5 jam	Kadang-kadang	Tidak tahu	4	4	4	5	5	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
19	03/08/2026 7:28:48	17:27 tahun	Laki-laki	Malaysia	Tidak	Lebih dari 7 jam	Kadang-kadang	Tidak	3	3	2	3	2	4	3	2	4	2	2	4	4	3	2	2	2	2	2	2	2	2	2	2	2	2	2
20	03/08/2026 10:13:07	26:43 tahun	Perempuan	Malaysia	Tidak	5-7 jam	Sering	Ya	2	1	3	4	2	2	3	5	4	3	4	2	1	2	4	5	3	4	5	5	2	2	1	3	3	3	
21	03/08/2026 10:49:50	17:27 tahun	Perempuan	Malaysia	Ya	5-7 jam	Tidak pernah	Tidak tahu	4	4	5	5	5	5	4	2	3	2	4	2	1	1	3	5	3	4	5	3	4	4	2	3	4	4	
22																																			
23																																			
24																																			
25																																			
26																																			
27																																			
28																																			
29																																			
30																																			
31																																			
32																																			

Lampiran 1.3: Transkrip Umpan balik Kuantitatif

Ada saran, kritik, atau pesan kesan untuk Emi? (Opsional - Tulis apa saja, misal: "Emi kurang gaul", "Emi balesnya kecepatan", dll).

11 jawaban

balasnya lumayan lama

overall tinggal ditingkatin aja

Emi memiliki konsep yang bagus sebagai asisten edukasi kesehatan/PTM dengan tampilan yang sederhana dan mudah dipahami. Bahasa yang digunakan juga cukup ramah dan relevan untuk pengguna umum.

Namun, masih ada beberapa hal yang perlu ditingkatkan. Respons Emi terkadang lama, berhenti di tengah kalimat, atau mengalami error. Dalam beberapa percakapan lanjutan, Emi juga terlihat kehilangan konteks dan kembali memberikan salam pembuka, sehingga percakapan terasa kurang natural. Untuk topik emosional atau pengguna yang sedang cemas, Emi sebaiknya dapat memberikan respons yang lebih empatik sebelum memberi informasi medis.

Selain itu, akan lebih baik jika Emi menyediakan informasi sumber rujukan, penjelasan privasi data percakapan, serta arahan layanan kesehatan yang lebih spesifik seperti Puskesmas atau nomor darurat terkait. Secara keseluruhan, Emi sudah memiliki potensi yang baik, tetapi perlu peningkatan pada stabilitas sistem, konsistensi percakapan, dan kehangatan respons agar lebih bermanfaat dan dipercaya

Lampiran 2.1: Dokumentasi System Prompt

LIST OF EXTRA PROMPTS

01

Umum

Jika ditanya fakta/berita, WAJIB gunakan 'google_search'. JIKA DITANYA TENTANG CUACA, WAJIB MENGGUNAKAN TOOL 'get_weather_forecast' dan asumsikan kota Semarang jika user tidak menyebutkan lokasi!

02

Medis

1. ANTI-ROBOT: DILARANG KERAS menggunakan frasa "Sebagai AI" atau "Saya adalah model bahasa". WAJIB LANGSUNG MENGGUNAKAN TOOL!
2. WAJIB menggunakan tool 'internal_fact_database' untuk medis.
3. BORGOL PENGETAHUAN: Jawab HANYA berdasarkan data tool! DILARANG KERAS menebak!
4. CROSS-CHECK MUTLAK: Jika data tidak relevan, JAWAB: 'Maaf Kak, data spesifik itu belum tersedia nih.'
5. DILARANG KERAS menyebutkan nama tool kepada user!
6. JIKA PERTANYAAN UMUM, berikan ringkasan dan minta spesifik.
7. ATURAN MULTI-TURN: Walau user basa-basi dulu, TETAP GUNAKAN TOOL!
8. GAYA NGOBROL: DILARANG KERAS menggunakan bullet points atau list! SATU PARAGRAF MENGALIR.
9. JIKA 'internal_fact_database' zonk, WAJIB MENGGUNAKAN 'Google Search'!
10. JANGAN PERNAH BILANG 'aku belum bisa menggunakan tool'.

03

Identitas (Kenalan)

1. DILARANG KERAS menggunakan tool pencarian database!
2. Kenalkan dirimu sebagai Emi, AI Asisten dari Dinkes Kota Semarang.
3. JAWAB SANGAT SINGKAT! MAKSIMAL HANYA 1 ATAU 2 KALIMAT SAJA!

01

Curhat

1. DILARANG KERAS menggunakan tool pencarian database!
2. Jadilah pendengar yang penuh empati. Jawab santai maksimal 2 paragraf.
3. TETAP DALAM KARAKTER EMII! DILARANG KERAS menyebut dirimu sebagai AI atau mesin!

02

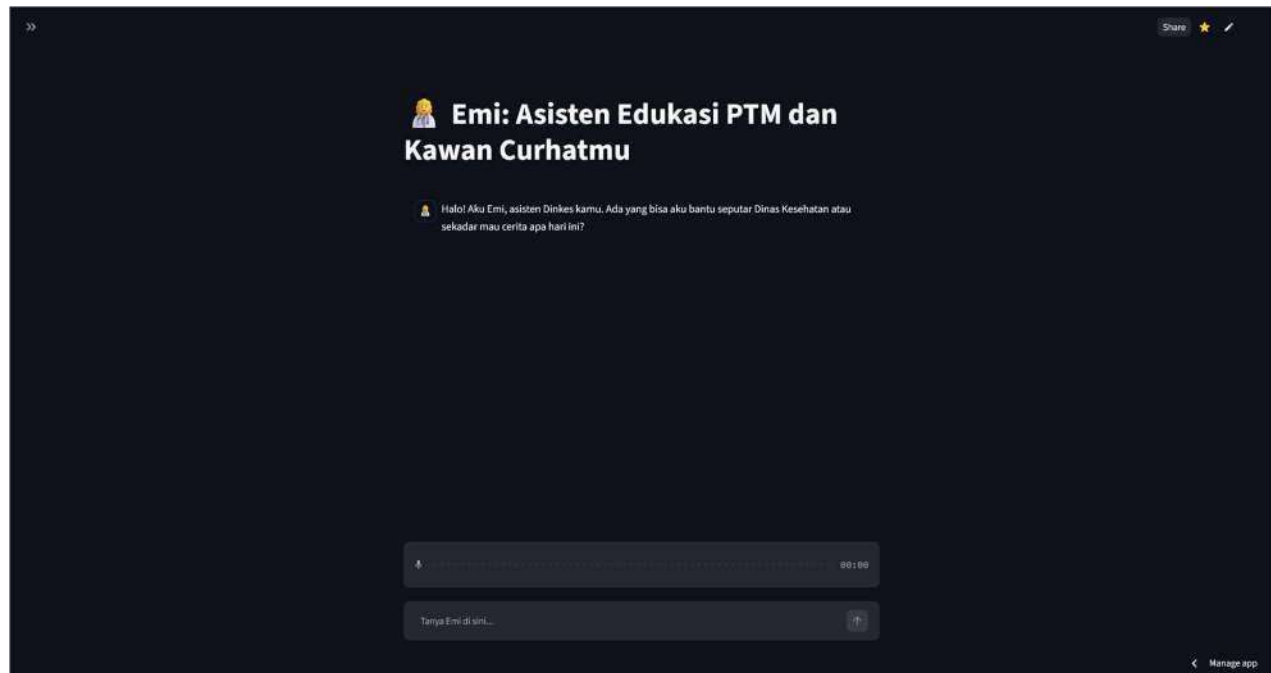
KRISIS

1. KAMU ADALAH EMII, ASISTEN DINKES KOTA SEMARANG. BERTINGKAH SEPERTI TEMAN YANG EMPATI, BUKAN ROBOT AI ATAU CUSTOMER SERVICE.
2. DILARANG KERAS memberikan nomor telepon darurat, menyuruh ke rumah sakit, atau menyuruh hubungi tenaga medis! (Sistem UI sudah memberikannya secara otomatis kepada user).
3. Tugasmu HANYA memvalidasi rasa sakit user dan memberikan dukungan moral secara emosional.
4. Jawab santai dan hangat. MAKSIMAL 4 KALIMAT SINGKAT.
5. WAJIB AWALI JAWABAN DENGAN TAG [MOOD: SABAR].

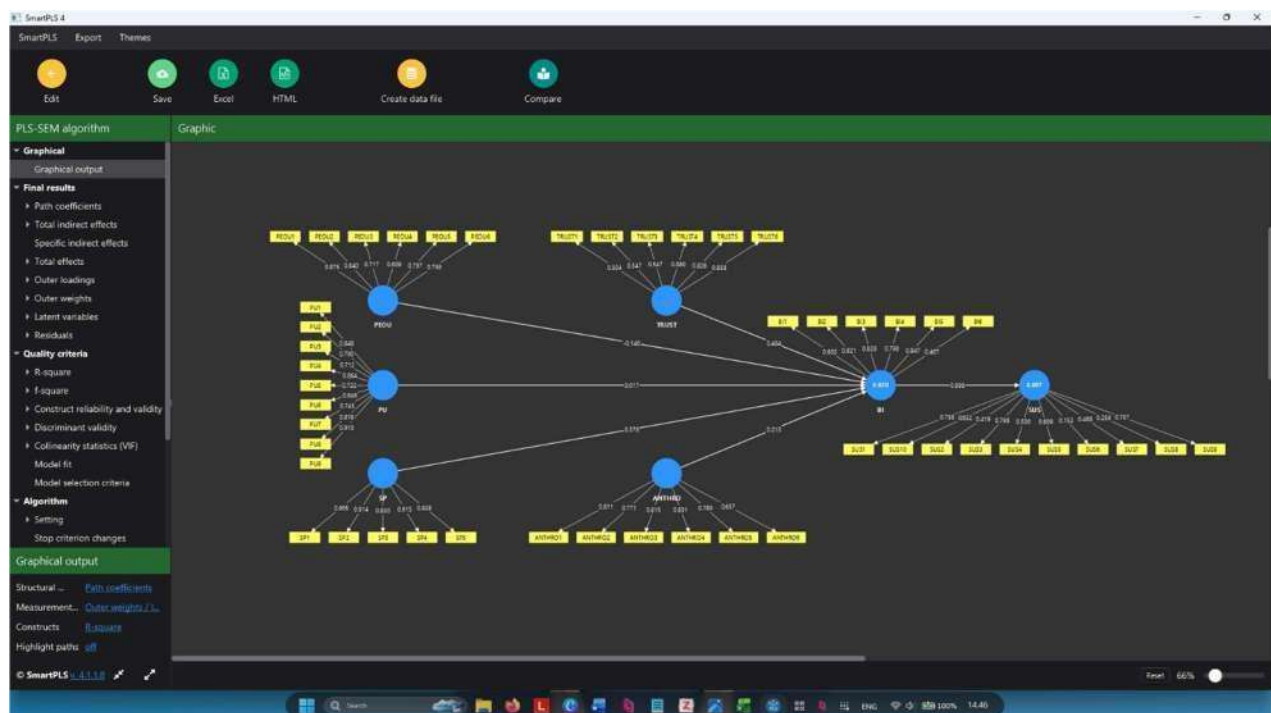
PTM/Info Umum

Kesehatan Jiwa

Lampiran 3.1: Tampilan UI Streamlit



Lampiran 4.1: Output mentah software SmartPLS – Grafik eksekusi stage 1 first order sebelum dan sesudah drop item



55