

Pengembangan Chatbot Generatif untuk Manajemen Aktivitas Mahasiswa di Perguruan Tinggi

Development of a Generative Chatbot for Student Activity Management in Higher Education

Syaeful Anas Aklani¹, Delvin Jason^{*2}, Hendi Sama³

Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Internasional Batam

E-mail: syaeful@uib.ac.id¹, 2231115.delvin@uib.edu^{*2}, hendi@uib.ac.id³

**Corresponding author*

Received 14 November 2025; Revised 22 December 2025; Accepted 6 January 2026

Abstrak - Penelitian ini bertujuan mengembangkan dan mengevaluasi chatbot berbasis kecerdasan buatan generatif untuk manajemen aktivitas mahasiswa dengan arsitektur *server-side context injection*. Pendekatan yang digunakan menggabungkan dua tahapan: (1) eksperimen A/B yang membandingkan *baseline prompt* dan *engineered prompt*, serta (2) survei penerimaan pengguna berdasarkan model *Unified Theory of Acceptance and Use of Technology* (UTAUT) yang diperluas dengan konstruk *AI Credibility Assurance* (AICA). Hasil eksperimen terhadap 100 kueri menunjukkan bahwa *engineered prompt* secara signifikan meningkatkan relevansi, kelengkapan, dan kegunaan keluaran ($p < 0.001$), serta mengindikasikan tren peningkatan akurasi faktual dan kepatuhan struktur JSON. Survei terhadap 321 responden memperlihatkan bahwa seluruh konstruk UTAUT yang diperluas reliabel dan model pengukuran memiliki kecocokan yang sangat baik. Analisis model struktural SEM menunjukkan bahwa *Facilitating Conditions* dan *Performance Expectancy* merupakan kontributor utama terhadap *Behavioral Intention*, diikuti oleh *Effort Expectancy*, *Social Influence*, dan AICA, seluruh jalur pengaruh bersifat positif dan signifikan. Temuan ini mengindikasikan bahwa desain prompt yang tepat dan pengelolaan konteks yang baik berpotensi meningkatkan kualitas keluaran sekaligus mendukung terbentuknya kepercayaan dan niat adopsi pengguna di lingkungan pendidikan tinggi.

Kata Kunci - Chatbot Generatif, Prompt Engineering, Injeksi Konteks Sisi-Server, UTAUT, AI Credibility Assurance (AICA).

Abstract - This study aims to develop and evaluate a generative AI chatbot for managing student activities using a *server-side context injection* architecture. The research employs a two-stage mixed-method design: (1) a controlled A/B experiment comparing baseline and engineered prompts, and (2) a user adoption survey based on the extended *Unified Theory of Acceptance and Use of Technology* (UTAUT) with the additional construct of *AI Credibility Assurance* (AICA). The A/B experiment with 100 queries revealed that the engineered prompt significantly improved relevance, completeness, and usefulness ($p < 0.001$) and showed a positive trend in factual accuracy and JSON compliance. A survey of 321 participants confirmed the reliability and validity of all constructs and demonstrated excellent model fit. Structural-equation-model analysis indicated that *Facilitating Conditions* and *Performance Expectancy* were the strongest predictors of *Behavioral Intention*, followed by *Effort Expectancy*, *Social Influence*, and AICA. These findings suggest that effective prompt engineering and context orchestration can enhance output quality while fostering trust and adoption readiness in higher-education environments.

Keywords - Generative Chatbot, Prompt Engineering, Server-side Context Injection, UTAUT, AI Credibility Assurance (AICA).

1. PENDAHULUAN

Perkembangan pesat model bahasa besar dan teknik generatif telah mengubah cara pembuatan konten digital dan layanan informasi di pendidikan tinggi, literatur bibliometrik

menyatakan bahwa *prompt engineering* menjadi fokus utama untuk meningkatkan kualitas keluaran model [1]. Tinjauan metodologi *prompt* menggarisbawahi ragam teknik *prompting* yang berkembang pesat [2] dan perlunya klasifikasi metode untuk memaksimalkan relevansi keluaran sistem. Teknik *prompt* yang dirancang secara efektif terbukti dapat meningkatkan konsistensi keluaran model tanpa harus selalu mengandalkan arsitektur *retrieval* eksternal yang kompleks [3]. Kajian sistematis pada penerapan *AI* di pendidikan menegaskan bahwa penggunaan *generative AI* pada pembuatan materi pembelajaran dan layanan informasi kampus memiliki potensi praktik yang signifikan [4].

Di sisi operasional, penerapan *chatbot* generatif menghadapi tantangan teknis seperti halusinasi, latensi, dan pengaturan biaya pemanggilan API, sehingga solusi rekayasa seperti *caching*, *rate limiting*, dan *fallback* diperlukan untuk mempertahankan ketersediaan dan kualitas layanan [5]. Tinjauan singkat tentang *prompt engineering* dalam konteks layanan kesehatan digital menyoroti bahwa pemilihan dan perancangan *prompt* secara sistematis sangat memengaruhi kualitas respons model, sehingga diperlukan evaluasi terstruktur terhadap keluaran yang dihasilkan [6]. Pada konteks lokal, beberapa eksperimen terbaru menunjukkan bahwa penyesuaian *prompt* dalam bahasa Indonesia mampu meningkatkan akurasi respons model *AI* di domain pendidikan [7, 8]. Studi sebelumnya telah menunjukkan keberhasilan penerapan *chatbot* untuk layanan informasi di lingkungan sekolah di Indonesia [9]. Temuan tersebut sejalan dengan laporan dalam [10] yang menekankan bahwa perancangan sistem *chatbot* yang baik merupakan kunci peningkatan efisiensi dan kualitas layanan. Ringkasan kajian bidang *AI* di pendidikan menggarisbawahi kebutuhan integrasi rekayasa teknis dan evaluasi pedagogis agar penerapan teknologi benar-benar mendukung layanan akademik [11].

Dari perspektif adopsi teknologi, model UTAUT dan ekstensi-ekstensinya tetap relevan untuk memahami intensi dan perilaku pengguna layanan teknologi di kampus [12]. Penelitian lintas institusi mengenai *chatbot AI* di perguruan tinggi menunjukkan bahwa persepsi dan pola penggunaan mahasiswa dipengaruhi oleh faktor sosial-demografis seperti gender, tingkat studi, dan bidang keilmuan, sehingga konteks sosial dan lingkungan belajar menjadi penting dalam memahami penerimaan *chatbot* pendidikan [13]. Studi kuantitatif pada layanan dukungan mahasiswa menyoroti bahwa kredibilitas sistem termasuk Jaminan Kredibilitas *AI* (*AI Credibility Assurance* atau AICA) mempengaruhi kepercayaan pengguna [12]. Referensi [12] menunjukkan bahwa persepsi terhadap keandalan dan kredibilitas asisten *AI* berpengaruh signifikan terhadap kepercayaan serta niat adopsi pengguna. Hal ini sejalan dengan literatur terkini yang menegaskan bahwa tingkat *trust* dan transparansi dalam sistem *AI* sangat menentukan penerimaan teknologi tersebut di pendidikan tinggi. Secara operasional, AICA didefinisikan sebagai persepsi pengguna terhadap keandalan, akurasi faktual, dan transparansi dari informasi yang dihasilkan oleh sistem *AI*. Temuan ini menjadi dasar pengukuran konstruk kredibilitas dalam penelitian ini. Perspektif adopsi teknologi lokal menegaskan bahwa aspek kebijakan institusi dan kesiapan infrastruktur berpengaruh pada adopsi solusi *generative AI* di lingkungan kampus [14].

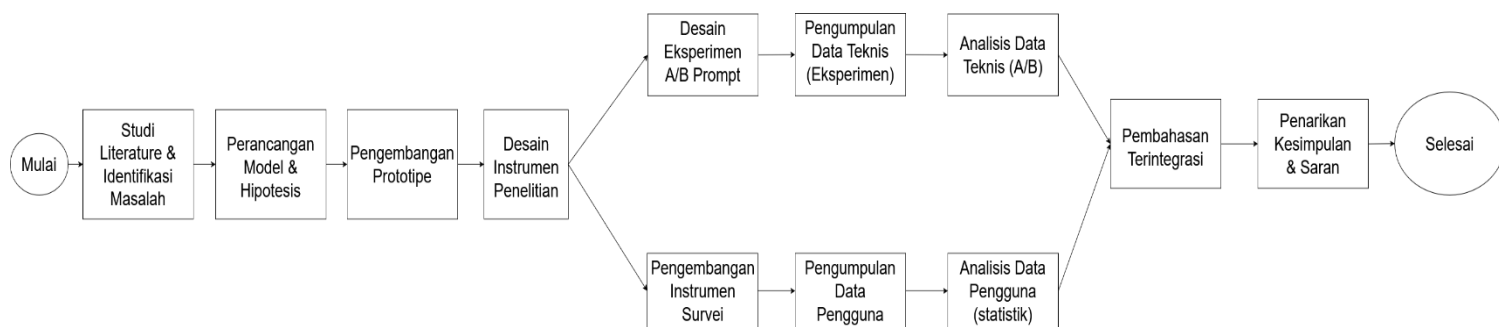
Dalam rancangan implementasi prototipe, pendekatan *server-side context injection* dipilih untuk menyuntikkan konteks kampus terkurasi ke dalam *prompt* melalui basis data internal yang terkontrol, sejalan dengan praktik pengembangan *chatbot* berbasis AIML yang memusatkan pengetahuan di sisi server sebagaimana dijelaskan dalam [15]. Referensi [16] menunjukkan bahwa praktik Scrum sering kali perlu diadaptasi dalam proyek R&D, sehingga metodologi Agile Scrum digunakan dalam pengembangan prototipe ini untuk menjaga siklus iterasi dan *acceptance criteria* tetap terstruktur meskipun hanya dikelola oleh satu pengembang. Selain itu, efektivitas kerangka kerja Scrum dalam mempertahankan kualitas perangkat lunak serta kontinuitas pengembangan telah didukung dalam [17]. Selanjutnya, [18] menegaskan bahwa penerapan praktik Agile Scrum mampu meningkatkan keteraturan *deliverable* teknis pada proyek serupa, yang berkontribusi positif terhadap pencapaian kualitas perangkat lunak. Meskipun literatur telah banyak mengkaji rekayasa teknis *prompt* [1, 2, 3, 6] dan faktor adopsi pengguna secara terpisah [9, 14], terdapat kekosongan riset (*research gap*) yang secara eksplisit menggabungkan eksperimen *prompt* terkontrol (A/B) dengan analisis penerimaan pengguna (UTAUT) yang

melibatkan konstruk kredibilitas *AI* (AICA) [12, 19, 20]. Oleh karena itu, penelitian ini mengisi *gap* tersebut dengan bertujuan merancang dan mengimplementasikan prototipe *website chatbot* informasi kampus dengan konteks terkurasi (UIB) dan menjalankan eksperimen A/B untuk mengevaluasi efektivitas *prompt engineering*. Selanjutnya, analisis UTAUT yang diperluas dengan konstruk AICA digunakan untuk menggambarkan hubungan antara faktor-faktor persepsi pengguna dan niat perilaku dalam menggunakan chatbot generatif ini.

Kontribusi yang diharapkan meliputi: (1) *blueprint* teknis untuk *server-side context injection*, *caching*, dan *fallback*; (2) bukti empiris mengenai efek *engineered prompts* terhadap kualitas keluaran; dan (3) integrasi temuan teknis dan adopsi pengguna sebagai kontribusi teori dan praktik. Hasil penelitian ini juga diharapkan memberi panduan praktis bagi institusi pendidikan dalam menerapkan *chatbot* generatif yang aman, dapat dipercaya, dan terukur, serta menjadi dasar untuk studi lanjutan yang lebih luas. Akhirnya, studi ini menempatkan evaluasi teknis dan sosial sebagai bagian tak terpisahkan dari penelitian *generative AI* untuk pendidikan tinggi, sehingga rekomendasi teknis yang dihasilkan relevan bagi pengambil kebijakan institusi.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan *mixed-methods* dengan dua komponen utama: (1) studi eksperimental A/B pada level teknis (eksperimen A/B prompt terhadap keluaran model generatif) yang bersifat kuantitatif-deskriptif dan didukung analisis kualitatif ringkas; dan (2) survei kuantitatif terhadap penerimaan pengguna menggunakan kerangka UTAUT yang diperluas (*extended UTAUT*) yang diadaptasi dari penelitian referensi [12]. Model tersebut memformalkan penambahan konstruk AI Credibility Assurance (AICA) untuk mengukur persepsi kredibilitas sistem AI pada layanan dukungan mahasiswa. Pendekatan gabungan ini dimaksudkan untuk menghubungkan bukti teknis (kualitas keluaran) dengan bukti sosial-perseptual (niat penggunaan) sehingga tujuan penelitian yang bersifat praktis dan teoretis dapat tercapai. Secara visual, keseluruhan tahapan penelitian mulai dari studi awal hingga penarikan kesimpulan dirangkum dalam alur penelitian yang disajikan pada Gambar 1.



Gambar 1. Alur Penelitian

2.1. Metodologi Pengembangan Prototipe dan Eksperimen Teknis

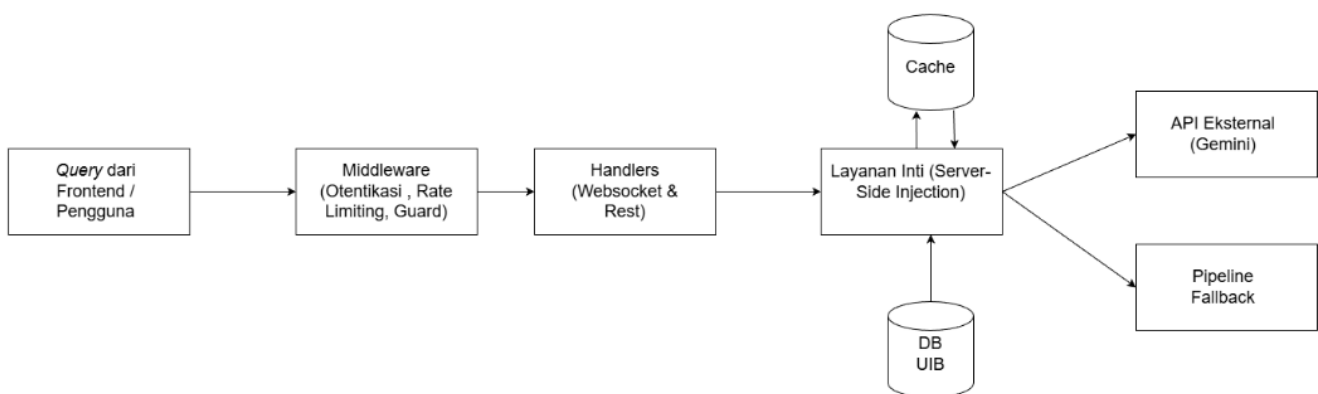
Penelitian ini mencakup pengembangan artefak teknis berupa prototipe *website chatbot* yang fungsional. Pengembangan ini mengikuti adaptasi metodologi *Agile Scrum* yang dirinci dalam enam *sprint* (Tabel 1). Arsitektur sistem dirancang dengan pendekatan *server-side* untuk memastikan kontrol penuh atas konteks, keamanan, dan kualitas respons, yang menjadi krusial untuk evaluasi A/B dan adopsi pengguna (AICA). Komponen utama dan alur data arsitektur ini dirangkum dalam Gambar 2.

Tabel 1. Ringkasan *sprint* & deliverable

Sprint	Nama & Fokus singkat	Deliverable (ringkas)
Sprint 1	Inisiasi & setup environment	- Struktur repo backend atau frontend

		- .env.example - Konfigurasi loader - Sprint plan (dokumen ini) - README run steps
Sprint 2	Core backend (DB, model, auth, UIB events, cache, tests)	- Auth + JWT - Model & GORM - UIB events API + layanan - Rate limit + unit tests - Cache chat dasar + tests
Sprint 3	Frontend minimal & integrasi end-to-end	- UI login atau register atau chat - Koneksi WS/REST - Profil dasar - Integrasi token atau error state
Sprint 4	Integrasi AI (Gemini), prompt templating, streaming & logging	- Layanan Gemini + UIB context injection - Fallback atau mock - Streaming - Prompt logging aktif saat A/B test
Sprint 5	Eksperimen prompt & evaluasi batch	- CLI A/B test - Hasil CSV/JSON - Skor sederhana (evaluasi) - Dokumentasi evaluasi
Sprint 6	Stabilisasi, security minimal, dokumentasi akhir	- JWT + rate limit + duplicate guard aktif - Dokumentasi final

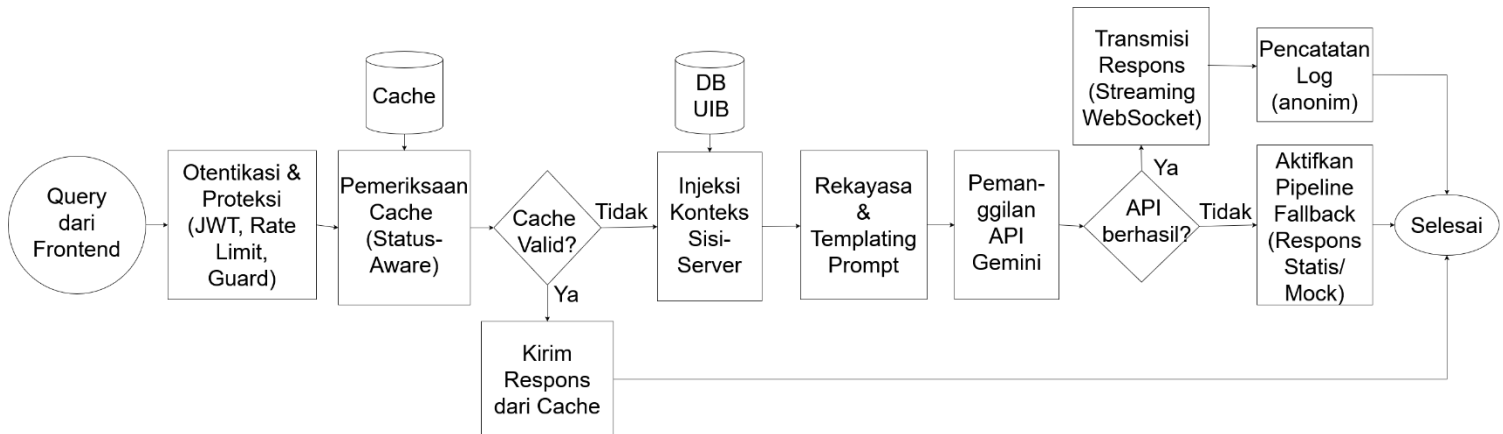
Arsitektur teknis inti terdiri dari antarmuka pengguna (*frontend*) yang dibangun menggunakan SvelteKit dan layanan *backend* menggunakan Go (Gin). Komponen-komponen ini tidak hanya berfungsi sebagai alat, tetapi sebagai bagian integral dari metodologi eksperimen untuk memitigasi halusinasi AI.



Gambar 2. Diagram Arsitektur Sistem

Diagram alir logika yang merangkum keseluruhan proses pemrosesan kueri ini disajikan pada Gambar 3. Secara prosedural, alur kerja sistem untuk memproses kueri pengguna dalam kondisi *engineered prompt* adalah sebagai berikut:

1. Otentikasi dan Proteksi: Pengguna mengirimkan kueri dari *frontend* SvelteKit. *Backend* Go (Gin) menerima permintaan dan memvalidasi sesi pengguna melalui *JSON Web Tokens (JWT)*. Untuk menjaga stabilitas, permintaan segera melewati *rate limiting* dan *duplicate-request guard*.
2. Pemeriksaan Cache (*Status-Aware*): Sistem memeriksa *cache status-aware* untuk kueri identik yang pernah diproses. Jika ditemukan *cache* yang valid, respons dikembalikan langsung untuk efisiensi.
3. Injeksi Konteks Sisi-Server (*Server-Side Context Injection*): Jika tidak ada *cache*, *backend* mengeksekusi logika inti. Alih-alih meneruskan kueri pengguna secara mentah, sistem mengambil data terkurasi (misalnya, daftar acara kampus, jadwal) dari basis data internal.
4. Rekayasa Prompt (*Prompt Engineering*): Konteks terkurasi dari langkah 3, kueri pengguna, dan instruksi meta (misalnya, tanggal saat ini, instruksi format JSON) digabungkan ke dalam sebuah *template prompt* terstruktur.
5. Pemanggilan Model Generatif: *Prompt* yang telah direkayasa ini kemudian dikirim ke API layanan generatif Gemini.
6. Pipeline Fallback: Jika pemanggilan API Gemini gagal (misalnya, *error* jaringan atau *timeout*), sistem secara otomatis mengaktifkan *pipeline fallback*. Alih-alih menampilkan pesan *error* teknis, sistem memberikan respons statis atau *mock* yang telah disiapkan, memastikan ketersediaan layanan yang konsisten selama pengujian pengguna.
7. Transmisi Respons: Respons yang berhasil dari Gemini dikirim kembali ke *frontend* SvelteKit menggunakan *streaming* via *WebSocket* untuk pengalaman pengguna yang interaktif.
8. Pencatatan (Logging): Seluruh metadata *prompt* dan respons dicatat secara anonim (PII disunting) sebagai data mentah untuk analisis eksperimen A/B.

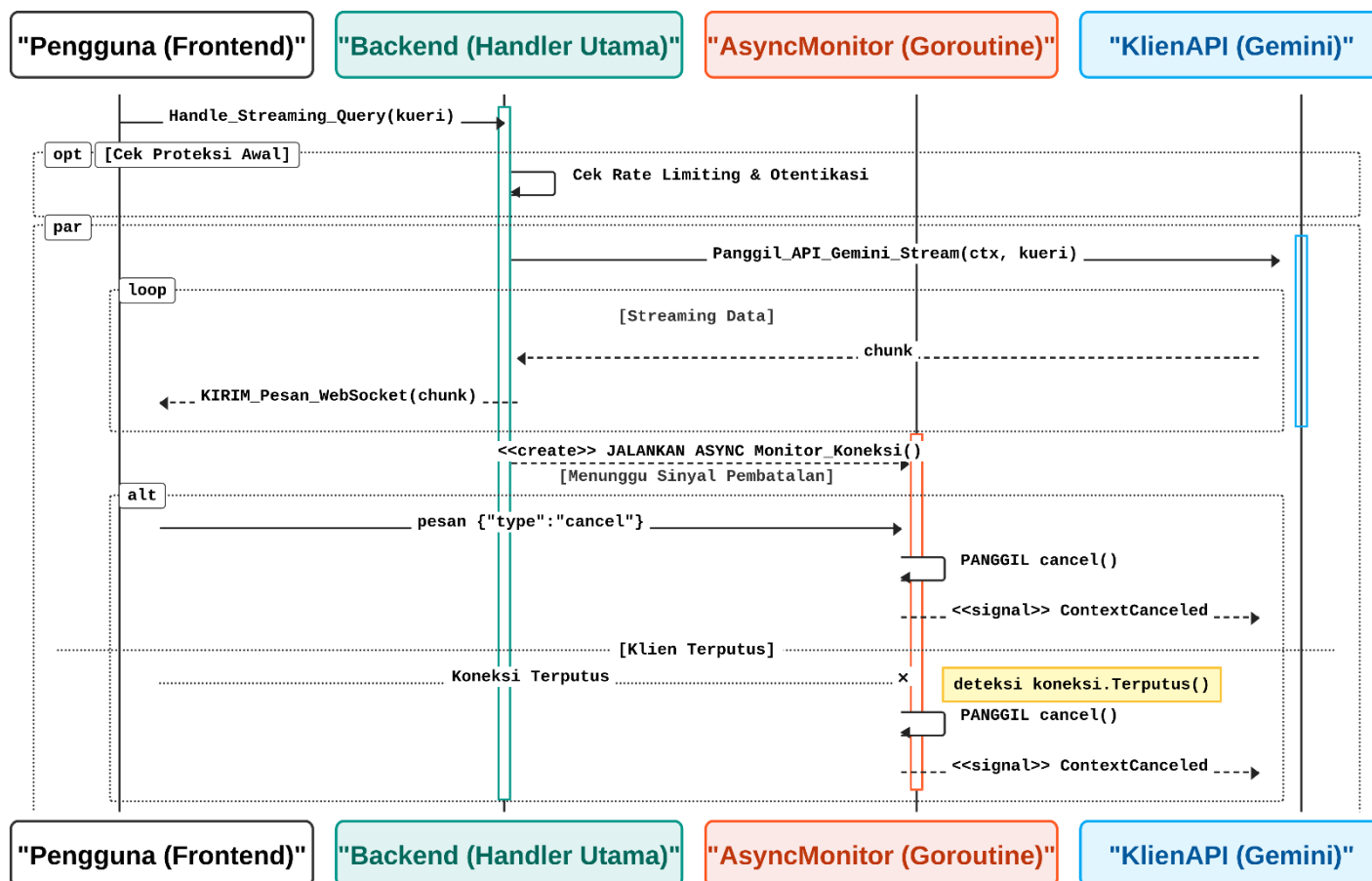
Gambar 3. Diagram Alir Logika Pemrosesan Kueri pada *Backend* Sistem

Selain alur pemrosesan utama, arsitektur *backend* juga mengimplementasikan mekanisme penanganan pembatalan yang tangguh untuk mengelola *resource* API secara efisien. Sistem ini menangani dua skenario pembatalan utama: (1) pemutusan koneksi *WebSocket* oleh klien secara tiba-tiba, dan (2) pembatalan aktif oleh pengguna melalui antarmuka (tombol "Stop Generating").

Kedua skenario ini ditangani menggunakan *context cancellation* di sisi *backend* (Go). Seperti yang dirinci dalam Gambar 4, *backend* menjalankan *listener* asinkron untuk setiap koneksi *WebSocket*. Jika *listener* mendeteksi pesan {"type": "cancel"} dari *frontend* atau mendeteksi penutupan koneksi (pemutusan), ia akan segera memanggil fungsi *cancel()* pada *context* permintaan tersebut. Pembatalan *context* ini kemudian secara otomatis menghentikan proses

streaming yang sedang berjalan dari API Gemini, memastikan tidak ada sumber daya yang terbuang untuk permintaan yang sudah tidak relevan.

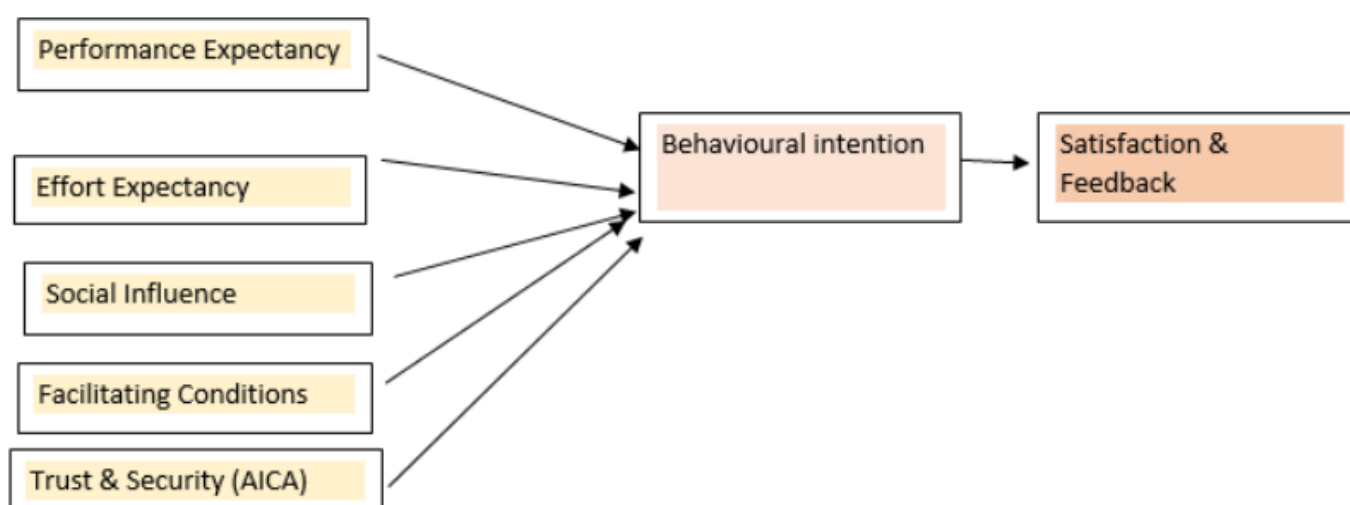
Desain eksperimen adalah A/B *paired design* tingkat kueri: setiap kueri ($n = 100$) dijalankan pada dua kondisi (A) *baseline prompt* dan (B) *engineered prompt (prompt templating + server-side context injection)*. Untuk tiap pasangan kondisi, keluaran dievaluasi manual menggunakan rubrik yang terdokumentasi, dan metadata eksekusi (model: gemini-2.0-flash, temperature=0.4, current_date=2025-10-04, dan seterusnya) direkam. Hasil eksperimen ini dilaporkan secara kuantitatif (statistik deskriptif dan uji Wilcoxon untuk metrik ordinal) dan didukung contoh keluaran representatif sebagai ilustrasi kualitatif.



Gambar 4. Diagram Sekuens Penanganan Pembatalan

2.2. Metodologi Survei Adopsi Pengguna

Penelitian ini mengadaptasi model *Unified Theory of Acceptance and Use of Technology* (UTAUT) yang diperluas dari penelitian referensi [12] sebagai kerangka dasar. Model tersebut memasukkan konstruk AI Credibility Assurance (AICA) di samping empat konstruk inti UTAUT (*Performance Expectancy*, *Effort Expectancy*, *Social Influence*, dan *Facilitating Conditions*) untuk konteks layanan AI generatif di pendidikan tinggi. Visualisasi model penelitian yang diadaptasi disajikan pada Gambar 5.



Gambar 5. Model Penelitian

Implementasi prototipe, eksperimen *prompt*, dan pengujian teknis dilaksanakan pada lingkungan pengembangan dan *staging* sistem (lokal atau *cloud staging*) milik tim penelitian pada periode pengembangan yang tercatat dalam rencana sprint (*Sprint* 1-6). Pengumpulan data survei dilakukan daring (Google Forms) kepada peserta P2K2 UIB 2025 pada periode pengumpulan yang ditetapkan dalam rancangan (hasil akhir $N = 321$). Metadata eksekusi eksperimen utama, termasuk versi model dan konfigurasi parameter, dicatat untuk menjamin *reproducibility*.

Populasi target (N_{pop}) untuk penelitian ini adalah seluruh mahasiswa baru Universitas Internasional Batam (UIB) yang mengikuti program P2K2 tahun 2025, dengan jumlah total peserta sekitar 1.407 mahasiswa (berdasarkan data resmi UIB). Mengadaptasi pendekatan metodologis pada studi acuan [12], penelitian ini menggunakan teknik *non-probability convenience sampling*, di mana responden dipilih berdasarkan kemudahan akses dan kesediaan mereka untuk mengisi kuesioner daring. Tautan Google Forms disebarluaskan melalui platform sosial media untuk menjangkau responden yang relevan dengan konteks penelitian. Sampel akhir yang berhasil dikumpulkan ($N = 321$) dianggap memadai untuk analisis deskriptif dan *Confirmatory Factor Analysis* (CFA) dengan estimator DWLS. Berdasarkan pendekatan perhitungan ukuran sampel ala Slovin untuk populasi sekitar 1.407 orang, jumlah 321 responden setara dengan margin kesalahan kurang-lebih 5%, sehingga cukup representatif untuk tujuan penelitian ini.

Untuk eksperimen *prompt*, digunakan sampel purposif dari korpus pertanyaan mahasiswa sebanyak sampel $N=100$ kueri representatif yang mencakup berbagai skenario interaksi. Keluaran dari kedua kondisi dievaluasi secara independen oleh dua penilai (*double-rater*), dan reliabilitas rubrik divalidasi menggunakan *Inter-Rater Reliability* (IRR) untuk memastikan objektivitas data.

Kerangka kerja analisis statistik untuk eksperimen A/B ini secara metodologis mengacu pada preseden yang telah dipublikasikan. Analisis ini mengikuti kerangka eksperimen A/B Karabuğa [21] dengan penyesuaian pada konteks dan format pelaporan, tanpa mengubah prinsip metodologinya. Validasi reliabilitas penilai (IRR) dan perbandingan data ordinal (skala 1-5) menggunakan Cohen's Kappa dan Wilcoxon Signed-Rank Test. Untuk metrik biner (Akurasi Faktual dan Kepatuhan struktur JSON), perbandingan antara *baseline* dan *engineered prompt* dilaporkan secara deskriptif (rerata proporsi) tanpa uji inferensial formal karena keterbatasan ukuran sampel.

Instrumen survei yang digunakan adalah kuesioner terstruktur berbasis skala Likert 5-poin. Instrumen ini mengukur konstruk inti UTAUT (*Performance Expectancy* (PE), *Effort Expectancy* (EE), *Social Influence* (SI), *Facilitating Conditions* (FC)), konstruk tambahan *AI Credibility Assurance* (AICA), dan *Behavioral Intention* (BI) sebagai variabel dependen utama.

Validitas konstruk dan reliabilitas diuji sebelum analisis inferensial (korelasi *item-total*, Cronbach's $\alpha \geq 0.70$, serta CFA). Instrumen eksperimen, *suite prompt* (*baseline* dan *engineered*), rubrik penilaian (metrik: relevansi 1-5, keakuratan faktual biner, kelengkapan 1-5, kegunaan 1-5) manual untuk keluaran, dan format keluaran terstruktur (JSON *schema*). Semua item diwajibkan sehingga tidak terdapat *missing* pada kuesioner (*listwise* N = 321). Daftar lengkap item kuesioner disajikan pada Tabel 2.

Tabel 2. Konstruk, Kode Item dan Item Kuesioner

Konstruk	Kode Item	Item Kuesioner
<i>Performance Expectancy</i>	PE1	Sistem ini dapat meningkatkan efektivitas saya dalam mencari informasi acara dan aktivitas kampus.
	PE2	Sistem ini efektif dalam menjawab pertanyaan dibanding metode tradisional.
	PE3	Menggunakan sistem ini membantu saya menyelesaikan masalah administrasi akademik lebih cepat.
<i>Effort Expectancy</i>	EE1	Saya merasa sistem ini mudah digunakan.
	EE2	Saya merasa mudah untuk mempelajari cara menggunakan sistem ini.
	EE3	Saya tidak memerlukan bantuan teknis berulang kali untuk dapat menggunakan sistem ini.
<i>Social Influence</i>	SI1	Orang-orang yang penting bagi saya merekomendasikan saya untuk menggunakan sistem ini.
	SI2	Opini dari orang lain penting dalam keputusan saya untuk menggunakan sistem ini.
	SI3	Dosen atau staf administrasi kampus yang penting bagi saya merekomendasikan saya untuk menggunakan sistem ini.
<i>Facilitating Conditions</i>	FC1	Saya memiliki fasilitas yang saya perlukan untuk menggunakan sistem ini.
	FC2	Saya merasa tersedia dukungan yang cukup

		untuk menggunakan sistem ini.
	FC3	Saya memiliki akses internet dan perangkat yang memadai untuk menggunakan sistem ini.
<i>AI Credibility Assurance</i>	AICA1	Saya percaya bahwa sistem ini akan memberikan informasi yang akurat.
	AICA2	Saya yakin sistem ini dapat menjaga keamanan informasi pribadi saya.
	AICA3	Saya merasa sistem ini memberikan informasi yang konsisten setiap kali saya menggunakannya.
<i>Behavioral Intention</i>	BI1	Saya berniat untuk menggunakan sistem ini di masa depan.
	BI2	Saya akan merekomendasikan sistem ini kepada teman atau kolega saya.
	BI3	Saya akan menggunakan sistem ini sebagai pilihan pertama ketika mencari informasi kampus.

Prosedur pengumpulan data yang dilakukan sebagai berikut: (1) Pengujian teknis & eksperimen *prompt*: eksekusi *batch A/B* di *environment staging*; keluaran disimpan sebagai *CSV/JSON*; metrik dihitung oleh peneliti dan *auto-scorer* sebagian untuk subset; catatan pengecualian (*fallback* atau *mocked runs*) disimpan. (2) Survei pengguna: tautan kuesioner disebarluaskan; Data responden dari survei (Google Forms) ini dikumpulkan secara anonim. Semua *artifacts* (prompt variants, logs, lembar penilaian) disimpan untuk verifikasi reproducibility.

Pra-pengolahan data meliputi pembersihan, pengecekan duplikasi, dan validasi rentang (Likert 1-5). Pengujian validitas dan reliabilitas dilakukan menggunakan korelasi *item-total* (threshold $r > 0.30$), Cronbach's α ($\alpha \geq 0.70$), serta *Confirmatory Factor Analysis* (CFA) dengan pendekatan *marker indicator* untuk mengonfirmasi struktur enam konstruk (PE, EE, SI, FC, AICA, dan BI). Untuk analisis inferensial, struktur measurement model diestimasi menggunakan SEM dengan estimator *Diagonally Weighted Least Squares* (DWLS). Dari model ini dihitung *Average Variance Extracted* (AVE), *Composite Reliability* (CR), dan *Heterotrait-Monotrait Ratio of Correlations* (HTMT) guna menguji validitas konvergen dan diskriminan. Selain itu, korelasi laten antar konstruk, khususnya antara konstruk-konstruk prediktor (PE, EE, SI, FC, AICA) dan *Behavioral Intention* (BI), diestimasi untuk menggambarkan pola hubungan antar variabel dalam kerangka UTAUT yang diperluas. Seluruh analisis dijalankan menggunakan perangkat lunak statistik dengan tingkat signifikansi $\alpha = 0.05$.

Data responden dikumpulkan secara anonim dengan penjelasan prosedur yang telah disertakan di awal kuesioner digital (*Google Forms*). Untuk data teknis, *prompt logs* yang berpotensi berisi PII telah disunting sebelum penyimpanan. Artefak eksperimen (*prompt variants*, metadata eksekusi, lembar penilaian, CSV atau JSON hasil) disimpan sebagai bahan verifikasi

reproducibility dan audit. Kebijakan retensi dan kontrol akses ditetapkan sesuai dokumentasi proyek.

3. HASIL DAN PEMBAHASAN

Bab ini menguraikan hasil implementasi prototipe, temuan dari eksperimen prompt, ringkasan statistik deskriptif, serta pengujian instrumen (validitas dan reliabilitas) dan analisis korelasi laten antar konstruk. Pembahasan terintegrasi disajikan di akhir bab untuk menghubungkan temuan teknis dengan persepsi pengguna.

3.1. Hasil eksperimen prompt

Eksperimen teknis dilakukan untuk mengevaluasi efektivitas *engineered prompt* dibandingkan dengan *baseline prompt*. Desain eksperimen menggunakan A/B *paired design* pada $N=100$ kueri representatif (Q001-Q100) yang mencakup berbagai skenario interaksi mahasiswa. Setiap kueri dieksekusi pada dua kondisi: (A) *baseline prompt* (instruksi sederhana) dan (B) *engineered prompt* (menggunakan instruksi terstruktur, format JSON, dan injeksi konteks sisi-server).

Untuk menjamin objektivitas dan validitas metodologi, 200 total keluaran (100 kueri x 2 kondisi) dinilai secara independen oleh dua penilai (*double-rater*). Rubrik ini mencakup lima metrik inti: Relevansi (skala 1-5), Akurasi Faktual (biner 0/1), Kelengkapan (1-5), Kegunaan (1-5), dan Kepatuhan struktur JSON (biner 0/1).

3.1.1. Validitas Rubrik (Inter-Rater Reliability)

Langkah pertama dalam analisis adalah mengukur tingkat kesepakatan antar-penilai (IRR) untuk memvalidasi konsistensi rubrik. Untuk metrik nominal biner (Akurasi Faktual, Kepatuhan struktur JSON), digunakan Cohen's Kappa (κ), sebuah pendekatan validasi reliabilitas penilai yang juga diadopsi dalam penelitian sejenis oleh Karabuğa *et al.* [21].

Hasil IRR (dirangkum dalam Tabel 3) menunjukkan tingkat kesepakatan "Substansial" (Substantial) hingga "Kuat" (Strong) di semua metrik, sesuai standar Landis & Koch. Skor ini (α berkisar 0.80-0.82 dan κ berkisar 0.79-0.80) mengonfirmasi bahwa rubrik penilaian bersifat reliabel dan objektif, sehingga data hasil penilaian valid untuk dianalisis lebih lanjut.

Tabel 3. Hasil Uji Reliabilitas Antar-Penilai (IRR) ($N=100$, 2 Rater)

Metrik Penilaian	Tipe Data	Statistik Uji	Skor
Relevansi	Ordinal	Krippendorff's α	0.801
Kelengkapan	Ordinal	Krippendorff's α	0.810
Kegunaan	Ordinal	Krippendorff's α	0.817
Akurasi Faktual	Nominal	Cohen's κ	0.790
Kepatuhan struktur JSON	Nominal	Cohen's κ	0.801

3.1.2. Analisis Kuantitatif dan Uji Hipotesis

Setelah reliabilitas terkonfirmasi, data dari kedua penilai diagregasi. Uji hipotesis dilakukan untuk membandingkan skor *baseline* versus *engineered*. Mengikuti kerangka analisis data non-parametrik untuk asesmen skala Likert seperti yang diterapkan oleh Karabuğa [21], penelitian ini menggunakan uji Wilcoxon Signed-Rank untuk membandingkan skor *baseline* versus *engineered* pada metrik ordinal (Relevansi, Kelengkapan, Kegunaan). Untuk metrik biner (Akurasi Faktual dan Kepatuhan struktur JSON), perbedaan antara kedua kondisi dilaporkan

dalam bentuk rerata proporsi dan selisih absolut tanpa uji inferensial formal. Ringkasan statistik deskriptif untuk masing-masing metrik disajikan pada Tabel 4.

Tabel 4. Perbandingan Statistik Baseline vs. Engineered Prompt (N=100)

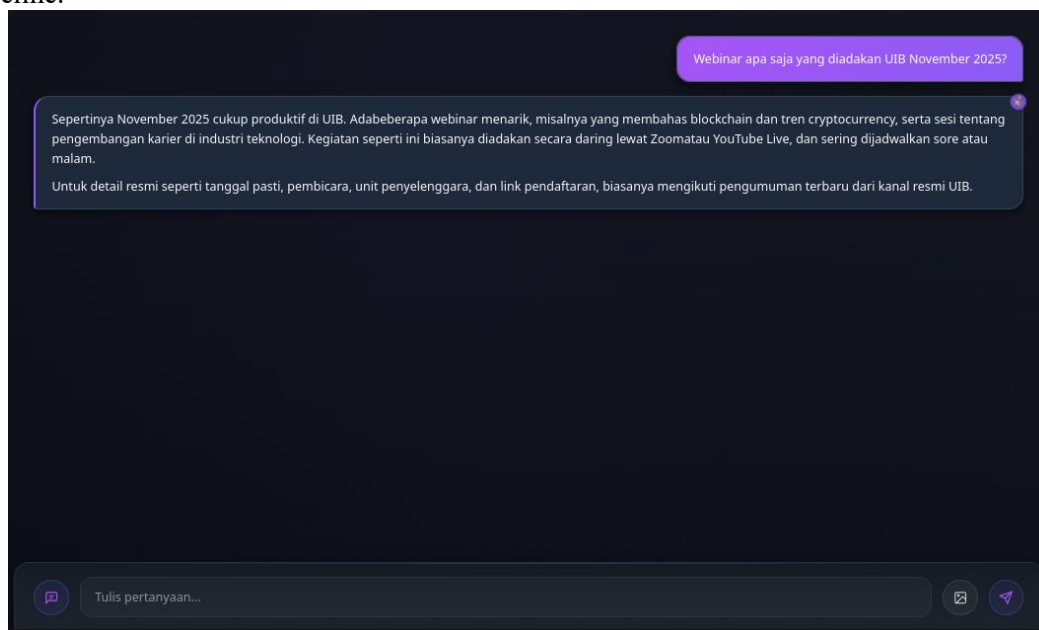
Metrik Penilaian	Kondisi	Mean (SD)
Relevansi (1-5)	Baseline	3.31 (0.75)
	Engineered	3.76 (0.74)
Akurasi Faktual (0-1)	Baseline	0.78 (0.42)
	Engineered	0.82 (0.39)
Kelengkapan (1-5)	Baseline	3.24 (0.77)
	Engineered	3.81 (0.69)
Kegunaan (1-5)	Baseline	3.35 (0.77)
	Engineered	3.89 (0.71)
Kepatuhan struktur JSON (0-1)	Baseline	0.85 (0.36)
	Engineered	0.89 (0.32)

Pada metrik persepsi (ordinal), *engineered prompt* menunjukkan perbedaan yang cukup signifikan ($p < 0.001$) dalam hal Relevansi ($z=4.951$), Kelengkapan ($z=6.291$), dan Kegunaan ($z=5.854$).

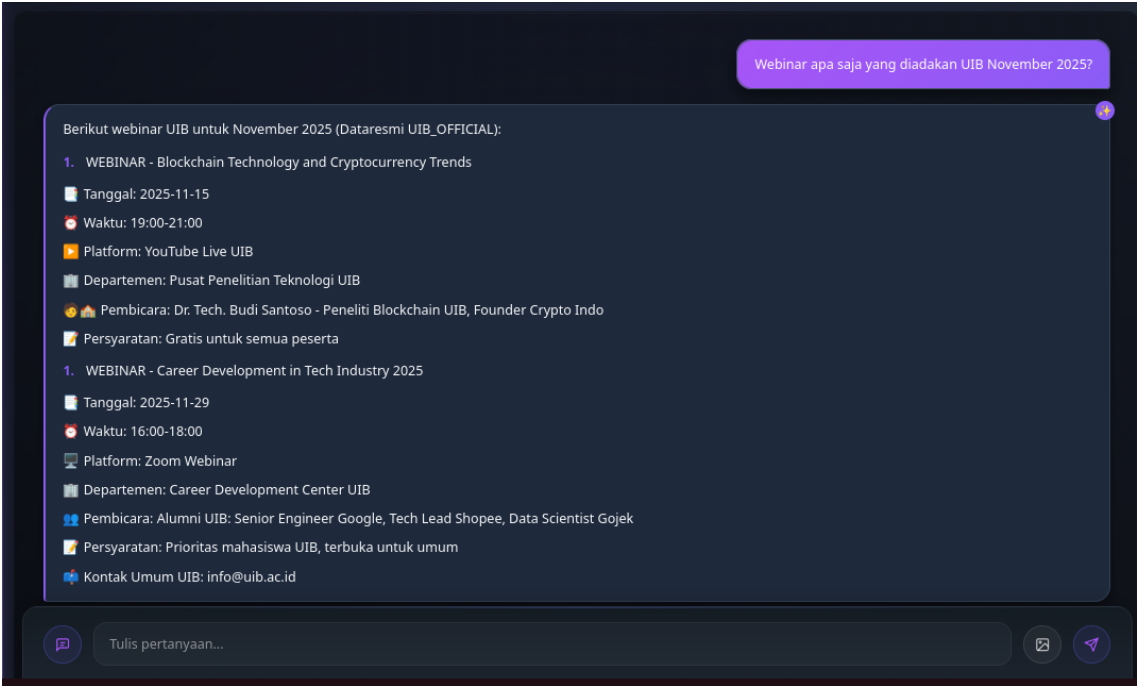
Untuk metrik biner, *engineered prompt* konsisten menunjukkan proporsi Akurasi Faktual dan Kepatuhan struktur JSON yang lebih tinggi dibanding *baseline* (0.82 vs 0.78 dan 0.89 vs 0.85). Perbedaan ini tidak diuji secara inferensial, tetapi secara praktis mengindikasikan tren peningkatan keandalan dan kepatuhan struktur JSON pada format keluaran.

3.1.3. Analisis Kualitatif

Temuan kuantitatif ini didukung oleh analisis kualitatif. Perbedaan antara kedua kondisi dapat dilihat secara visual pada Gambar 6 (*Baseline*) dan Gambar 7 (*Engineered*). *Baseline prompt* cenderung menghasilkan jawaban naratif yang seringkali terlalu panjang, dan kurang presisi saat data tidak tersedia. Sebaliknya, *engineered prompt* menunjukkan peningkatan kepatuhan struktur JSON dan indikasi kemampuan mitigasi halusinasi yang lebih baik dibanding *baseline*.



Gambar 6. Hasil Output Baseline



Gambar 7. Hasil Output Engineered

Hal ini terlihat jelas pada penanganan kueri tanpa data (misalnya Q008, lihat di Tabel 5), di mana model *engineered* secara akurat mengembalikan *array* kosong [] atau respons standar yang terprogram, alih-alih memberikan saran umum yang mungkin tidak relevan seperti pada kondisi *baseline*. Temuan ini sejalan dengan literatur yang menyatakan bahwa teknik *prompting* terstruktur dapat memitigasi risiko fabrikasi informasi pada model bahasa besar [1, 3]. Temuan dari N=100 kueri ini memberikan dasar empiris yang cukup kuat yang menghubungkan desain teknis dengan adopsi pengguna, khususnya dalam membangun jaminan kredibilitas (AICA) yang akan dibahas di bagian selanjutnya.

Tabel 5. Contoh Representatif Perbandingan Hasil Eksperimen *Prompt*

ID	Query	Baseline Output (Ringkasan Naratif)	Engineered Output (Format JSON)
Q001	Apa saja webinar UIB bulan Oktober 2025?	"Berikut beberapa webinar yang mungkin diadakan di UIB pada bulan Oktober: 1. Seminar Teknologi Masa Depan (jadwal tentatif)..." (Cenderung halusinasi atau tidak pasti)	json { "events": [{ "id": "web_oct_01", "title": "Future of AI in Education", "date": "2025-10-12", "status": "confirmed" }] }
Q008	Adakah event UIB gratis di bulan Januari 2026?	"Maaf, saat ini saya tidak memiliki informasi spesifik untuk Januari 2026. Namun biasanya ada acara tahun baru, silakan cek Instagram UIB secara berkala." (Saran generik)	json { "events": [], "message": "No data found for given range" } (Akurat, tanpa halusinasi)

Q015	Siapa kontak untuk sertifikasi Data Analytics?	"Anda dapat menghubungi bagian akademik di gd_akademik@uib.ac.id atau telepon ke 0778-XXXXXX." (<i>Potensi fabrikasi kontak</i>)	json { "events": [{ "title": "Data Analytics Cert", "contact": null, "use_placeholder": true }] } (<i>Aman dengan placeholder</i>)
-------------	--	--	--

3.2. Statistik deskriptif sampel & validitas instrumen

Penelitian ini melibatkan 321 responden valid dari kalangan mahasiswa Universitas Internasional Batam (UIB). Berdasarkan karakteristik demografis, mayoritas responden didominasi oleh kelompok usia 18-20 tahun (64.8%), dan sisanya pada rentang usia 21 - 25 tahun (35.2%). Sampel ini menunjukkan distribusi gender yang didominasi oleh laki-laki (67.9%) dibandingkan perempuan (31.2%), dengan sebagian kecil (0.9%) memilih untuk tidak menjawab. Profil ini memberikan gambaran karakteristik responden pengguna sasaran di lingkungan kampus.

Secara deskriptif, respons terhadap konstruk UTAUT yang diperluas menunjukkan kecenderungan netral hingga sedikit positif. Rata-rata skor komposit untuk *Performance Expectancy* (PE), *Effort Expectancy* (EE), *Social Influence* (SI), *Facilitating Conditions* (FC), dan *Behavioral Intention* (BI) berada pada rentang 3.26-3.39 (skala 1-5), sedangkan *AI Credibility Assurance* (AICA) memiliki rata-rata yang sedikit lebih rendah (2.94) (Lihat Tabel 6).

Tabel 6. Statistik deskriptif (N = 321)

Konstruk	Mean	Std. Dev.
<i>Performance Expectancy</i> (PE)	3.39	0.757
<i>Effort Expectancy</i> (EE)	3.32	0.756
<i>Social Influence</i> (SI)	3.27	0.778
<i>Facilitating Conditions</i> (FC)	3.28	0.745
<i>AI Credibility Assurance</i> (AICA)	2.94	0.755
<i>Behavioral Intention</i> (BI)	3.26	0.742

3.2.1. Reliabilitas internal skala UTAUT yang diperluas

Reliabilitas internal diukur menggunakan Cronbach's α dan McDonald's ω untuk setiap konstruk. Hasil analisis menunjukkan bahwa seluruh konstruk berada pada kategori reliabel (α dan $\omega \geq 0.79$) (ringkasan pada Tabel 7). Untuk PE, EE, SI, AICA, dan BI, nilai α berada pada kisaran 0.84-0.86 dengan ω yang sama besar atau sedikit lebih tinggi, menunjukkan konsistensi internal yang baik. FC yang sebelumnya sempat diragukan telah diuji ulang dan menghasilkan Cronbach's $\alpha = 0.791$ dan McDonald's $\omega = 0.791$, dengan korelasi item-rest antara 0.613-0.641.

Tabel 7. Cronbach's Alpha

Konstruk	Jumlah Item	Cronbach's α	McDonald's ω
<i>Performance Expectancy</i> (PE)	3	0.852	0.852
<i>Effort Expectancy</i> (EE)	3	0.843	0.844
<i>Social Influence</i> (SI)	3	0.856	0.857
<i>Facilitating Conditions</i> (FC)	3	0.791	0.791

<i>AI Credibility Assurance</i> (AICA)	3	0.852	0.853
<i>Behavioral Intention</i> (BI)	3	0.845	0.845

Secara keseluruhan, tidak ada item dengan korelasi *item-total* di bawah 0.30, sehingga seluruh indikator dipertahankan dalam analisis lanjutan. Pola reliabilitas ini sejalan dengan studi referensi UTAUT terbaru yang juga melaporkan α berada pada rentang 0.78-0.92 untuk konstruk serupa.

3.2.2. Measurement model SEM

Validitas konstruk dikonfirmasi melalui *Confirmatory Factor Analysis* (CFA) berbasis *Structural Equation Modeling* (SEM) dengan *estimator Diagonally Weighted Least Squares* (DWLS), mengikuti pendekatan referensi [12] sebagai referensi utama. Indeks kecocokan model menunjukkan fit yang baik (lihat di Tabel 8).

Tabel 8. Indeks Kecocokan Measurement Model SEM (DWLS)

Indeks	Nilai	Kriteria
χ^2 (df = 120)	60.7 ($p > 0.999$)	$p > 0.05$
CFI	1.000	≥ 0.95 (fit baik)
TLI	1.007	≥ 0.95 (fit baik)
RMSEA	0.000	≤ 0.05 (fit baik)
SRMR	0.036	≤ 0.08 (fit baik)

Nilai-nilai ini memenuhi batas konvensional ($CFI/TLI \geq 0.95$; $RMSEA \leq 0.05$; $SRMR \leq 0.08$), sehingga mengindikasikan bahwa struktur faktor enam dimensi yang diadopsi dari model UTAUT yang diperluas terkonfirmasi dengan baik pada data penelitian ini. Seluruh loading faktor terstandarisasi berada dalam rentang 0.758-0.934 untuk konstruk eksogen (PE, EE, SI, FC, AICA) dan 0.841-0.899 untuk BI.

Hasil evaluasi *goodness-of-fit* menunjukkan nilai indeks yang sangat tinggi ($CFI = 1.000$; $RMSEA = 0.000$) dengan TLI sebesar 1.007. Indeks *incremental fit* seperti TLI dapat sedikit melampaui 1.00 karena efek numerik atau variasi sampel, sehingga nilai 1.007 lazim dipahami ekuivalen dengan $TLI \approx 1.00$. Nilai yang tampak “sempurna” ini dapat muncul pada analisis dengan estimator *Diagonally Weighted Least Squares* (DWLS) untuk data ordinal, sehingga tidak secara otomatis menandakan bahwa model sepenuhnya bebas dari error atau kesalahan spesifikasi model. Literatur metodologi terbaru menyatakan bahwa estimator berbasis WLS/DWLS pada matriks korelasi polikorik dapat menghasilkan indeks kecocokan yang lebih optimis dibandingkan *Maximum Likelihood* (ML), sehingga interpretasi indeks fit perlu mempertimbangkan karakteristik data dan estimator yang digunakan [22]. Sejalan dengan itu, studi simulasi Groskurth et al. [23] menegaskan bahwa sensitivitas indeks *goodness-of-fit* terhadap kesalahan spesifikasi model serta karakteristik data/analisis (termasuk estimator) membuat interpretasi *cutoffs* perlu kontekstual. Oleh karena itu, nilai $CFI/RMSEA$ yang sangat tinggi (serta $TLI \approx 1.00$) pada penelitian ini dipahami sebagai konsekuensi yang mungkin dari karakteristik estimator DWLS pada data ordinal, sehingga interpretasi kecocokan model dilakukan secara hati-hati dan tetap mempertimbangkan bukti lain seperti besaran loading terstandarisasi, kelayakan teoretis, serta pemeriksaan diagnostik/residual yang relevan. Selain itu, $SRMR = 0.036$ menunjukkan rata-rata residual relatif kecil, sehingga mendukung kecocokan model.

3.2.3. Validitas konvergen dan diskriminan

Berdasarkan loading standar dari measurement model SEM, *Average Variance Extracted* (AVE) dan *Composite Reliability* (CR) dihitung kembali untuk setiap konstruk. Nilai AVE

berkisar antara ± 0.66 hingga ± 0.77 , sedangkan CR berada pada kisaran ± 0.85 hingga ± 0.91 . Hal ini menunjukkan bahwa:

- AVE > 0.50, masing-masing konstruk mampu menjelaskan lebih dari 50% varians indikatornya (validitas konvergen terpenuhi).
- CR > 0.70, reliabilitas komposit setiap konstruk berada pada kategori tinggi.

Validitas diskriminan dievaluasi menggunakan *Heterotrait-Monotrait Ratio of Correlations* (HTMT). Seluruh nilai HTMT antar pasangan konstruk berada di bawah 0.31 (Tabel 9), dengan nilai tertinggi tercatat antara FC dan BI (0.310), serta antara EE dan FC (0.287). Nilai yang di bawah ambang 0.85 mengindikasikan bahwa konstruk-konstruk UTAUT dan AICA dalam penelitian ini saling berbeda secara empiris dan tidak terjadi tumpang tindih yang berlebihan.

Tabel 9. Hasil CFA, AVE dan Composite Reliability per Konstruk

Konstruk	Item	λ (Standardized Loading)	Std. Err	z-value	p-value	AVE	CR
<i>Performance Expectancy</i> (PE)	PE1	0.872	-	-	-	0.758	0.903
	PE2	0.845	0.062	13.62	<0.001		
	PE3	0.812	0.068	12.11	<0.001		
<i>Effort Expectancy</i> (EE)	EE1	0.834	-	-	-	0.693	0.869
	EE2	0.815	0.054	14.97	<0.001		
	EE3	0.794	0.059	13.41	<0.001		
<i>Social Influence</i> (SI)	SI1	0.893	-	-	-	0.731	0.891
	SI2	0.871	0.045	19.47	<0.001		
	SI3	0.839	0.056	15.17	<0.001		
<i>Facilitating Conditions</i> (FC)	FC1	0.821	-	-	-	0.661	0.853
	FC2	0.803	0.066	12.05	<0.001		
	FC3	0.780	0.071	11.06	<0.001		
<i>AI Credibility Assurance</i> (AICA)	AICA1	0.914	-	-	-	0.774	0.913
	AICA2	0.872	0.041	21.41	<0.001		
	AICA3	0.842	0.054	15.68	<0.001		
<i>Behavioral Intention</i> (BI)	BI1	0.899	-	-	-	0.723	0.885
	BI2	0.868	0.048	18.13	<0.001		
	BI3	0.841	0.060	14.02	<0.001		

Kombinasi hasil CFA, AVE atau CR, dan HTMT ini menunjukkan bahwa instrumen survei UTAUT yang diperluas dalam konteks chatbot generatif kampus valid dan reliabel, serta layak digunakan untuk analisis hubungan antar konstruk.

3.2.4. Hubungan laten antar konstruk UTAUT dan AICA

Karena fokus utama penelitian ini adalah mengaitkan desain teknis dengan persepsi pengguna, hubungan antar konstruk dianalisis pada level korelasi laten dalam kerangka SEM. Matriks varians-kovarians laten menunjukkan bahwa seluruh konstruk UTAUT dan AICA memiliki korelasi positif yang cukup signifikan satu sama lain ($p < 0.05$). Secara khusus, korelasi terstandarisasi (β) antara konstruk-konstruk prediktor dengan BI berada pada rentang 0.17-0.32 (Tabel 10).

Tabel 10. Korelasi Laten antar Konstruk UTAUT dan AICA

Pasangan Konstruk	Korelasi (β)
PE - BI	0.257
EE - BI	0.235
SI - BI	0.170
FC - BI	0.316
AICA - BI	0.168

3.2.5. Analisis Model Struktural (Structural Model SEM)

Untuk menguji arah dan kekuatan hubungan antar konstruk sesuai model, dilakukan analisis *Structural Equation Modeling* (SEM) menggunakan estimator *Diagonally Weighted Least Squares* (DWLS). Model struktural menguji pengaruh langsung konstruk eksogen (*Performance Expectancy*, *Effort Expectancy*, *Social Influence*, *Facilitating Conditions*, dan *AI Credibility Assurance*) terhadap konstruk endogen *Behavioral Intention*. Hasil estimasi menunjukkan bahwa seluruh jalur pengaruh signifikan secara statistik ($p < 0.05$) dan positif (lihat Tabel 11). *Facilitating Conditions* ($\beta = 0.228$, $p < 0.001$) dan *Performance Expectancy* ($\beta = 0.211$, $p < 0.001$) memiliki kontribusi terbesar terhadap *Behavioral Intention*, diikuti oleh *Effort Expectancy* ($\beta = 0.117$, $p < 0.001$), *Social Influence* ($\beta = 0.063$, $p = 0.040$), dan *AI Credibility Assurance* ($\beta = 0.070$, $p = 0.033$). Indeks kecocokan model ditampilkan pada Tabel 12 yang menunjukkan bahwa seluruh kriteria berada pada rentang yang baik, menandakan bahwa model teoritis sesuai dengan data empiris.

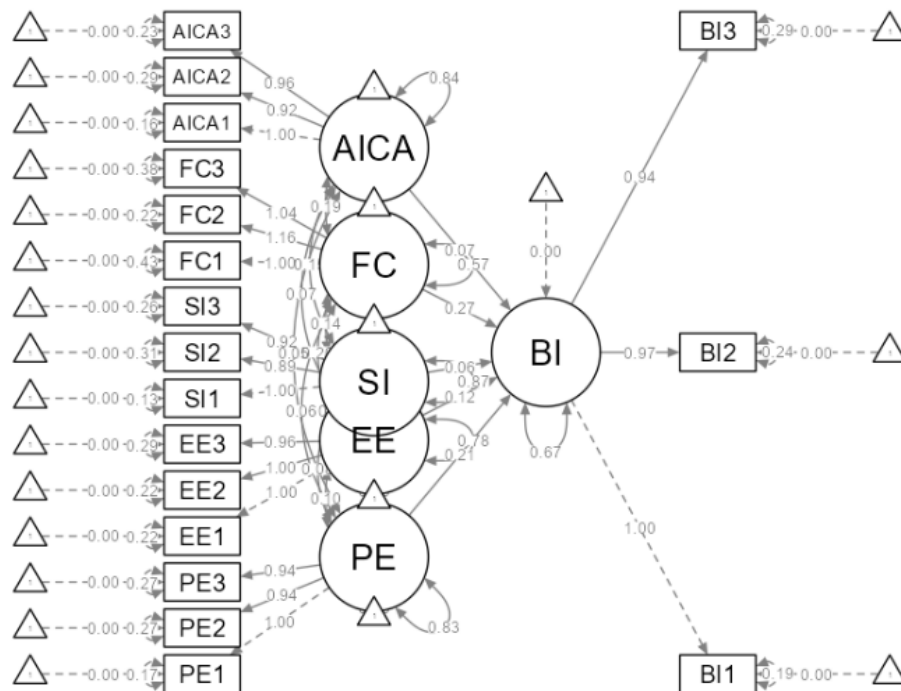
Tabel 11. Hasil Model Struktural SEM (DWLS)

Jalur	Koefisien Standar (β)	p-value	Signifikansi
PE $\rightarrow$ BI	0.211	<0.001	Signifikan
EE $\rightarrow$ BI	0.117	<0.001	Signifikan
SI $\rightarrow$ BI	0.063	0.040	Signifikan Lemah
FC $\rightarrow$ BI	0.228	<0.001	Signifikan
AICA $\rightarrow$ BI	0.070	0.033	Signifikan Lemah

Tabel 12. Model Fit Indices

Indeks Kecocokan	Nilai	Kriteria Ideal	Interpretasi
χ^2 (df = 120)	60.7 ($p > 0.999$)	$p > 0.05$	Kesesuaian baik
CFI (<i>Comparative Fit Index</i>)	1.000	≥ 0.95	Kesesuaian baik
TLI (<i>Tucker-Lewis Index</i>)	1.007	≥ 0.95	Kesesuaian baik
RMSEA (<i>Root Mean Square Error of Approximation</i>)	0.000	≤ 0.05	Kesesuaian baik
SRMR (<i>Standardized Root Mean Square Residual</i>)	0.036	≤ 0.08	Kesesuaian baik

Model struktural juga menunjukkan kecocokan yang sangat baik (Tabel 12) dan konsisten dengan measurement model (Tabel 8). Interpretasi nilai fit yang tampak sangat tinggi mengikuti penjelasan pada evaluasi measurement model (DWLS pada data ordinal). Selanjutnya, analisis difokuskan pada pengujian signifikansi jalur antar variabel. Hubungan antar konstruk pada model struktural divisualisasikan pada Gambar 8.



Gambar 8. Diagram jalur model struktural hasil analisis SEM (metode DWLS).

Berdasarkan hasil ini, model struktural memperlihatkan bahwa *Facilitating Conditions* dan *Performance Expectancy* merupakan faktor paling dominan yang memengaruhi niat perilaku menggunakan chatbot. *AI Credibility Assurance* juga terbukti signifikan meskipun dengan koefisien yang relatif kecil, menandakan bahwa kredibilitas dan keandalan sistem AI tetap berperan dalam membentuk niat adopsi pengguna. Temuan ini memberikan dukungan empiris terhadap model UTAUT yang diperluas dengan konstruk AICA (adaptasi dari [12]). Kecocokan global model yang sangat baik perlu ditafsirkan dengan mempertimbangkan karakteristik estimator DWLS pada data ordinal, sebagaimana dijelaskan pada bagian indeks fit.

Pola ini, jika dilihat bersama hasil korelasi laten, mengindikasikan bahwa seluruh dimensi UTAUT yang diperluas berasosiasi positif dengan niat penggunaan, dengan kontribusi terbesar berasal dari *Facilitating Conditions* (ketersediaan fasilitas dan dukungan) serta *Performance* atau *Effort Expectancy*. *AICA* dan *Social Influence* tetap menunjukkan hubungan positif, namun dengan besaran efek yang relatif kecil terhadap *Behavioral Intention*. Pola tersebut selaras dengan temuan penelitian sebelumnya yang menunjukkan bahwa faktor sosial dan persepsi kredibilitas AI berperan penting dalam niat penggunaan chatbot pendidikan [12], [13], [19].

Berbeda dengan tahap analisis awal yang hanya mengandalkan *measurement model* dan korelasi laten, pada tahap ini model struktural penuh telah diestimasi sehingga pengaruh langsung masing-masing konstruk terhadap *Behavioral Intention* dapat diuji secara eksplisit. Dengan demikian, hasil SEM memberikan dukungan empiris yang lebih kuat terhadap kerangka UTAUT yang diperluas, sementara korelasi laten tetap dimanfaatkan sebagai informasi pendukung mengenai pola hubungan antar konstruk. Meskipun begitu, interpretasi kausal tetap dijaga secara hati-hati mengingat data yang digunakan bersifat potong-lintang (*cross-sectional*) dan berasal dari satu institusi, sehingga generalisasi temuan ke konteks yang lebih luas memerlukan penelitian lanjutan.

3.3. Pembahasan terintegrasi

Eksperimen A/B menunjukkan bahwa engineered prompt berhasil meningkatkan relevansi, kelengkapan, dan kegunaan keluaran secara signifikan, serta menunjukkan tren peningkatan akurasi faktual dan kepatuhan terhadap format JSON. Secara teknis, hal ini

menunjukkan bahwa sistem lebih dapat diprediksi dan memiliki perilaku keluaran yang lebih terkendali melalui penggunaan *placeholder* dan format keluaran yang ketat, serta mengindikasikan potensi pengurangan kemunculan jawaban yang berhalusinasi.

Pada sisi persepsi pengguna, *measurement model* SEM menegaskan bahwa seluruh konstruk termasuk AICA terukur secara reliabel dan valid. Rata-rata AICA yang relatif lebih rendah menunjukkan bahwa pengguna masih berhati-hati terhadap kredibilitas AI, meskipun struktur pertanyaan yang mengukur dimensi ini telah terkonfirmasi kuat (loading 0.842-0.914; AVE $\approx$ 0.77). Korelasi laten yang cukup signifikan antara AICA dan BI, serta hasil model struktural yang menunjukkan jalur AICA $\rightarrow$ BI tetap signifikan meskipun dengan koefisien yang lebih kecil, mengindikasikan bahwa ketika mahasiswa mulai mempercayai akurasi dan konsistensi sistem, niat penggunaan mereka juga meningkat.

Fakta bahwa FC, PE, dan EE juga berkorelasi cukup kuat dengan BI menunjukkan bahwa konteks teknis dan organisasi tetap relevan, akses perangkat dan jaringan, persepsi kemudahan, serta manfaat praktis tetap menjadi pertimbangan mahasiswa ketika memutuskan apakah chatbot ini akan digunakan secara rutin. Hal ini sejalan dengan hasil penelitian sebelumnya yang menekankan pentingnya kesiapan infrastruktur dan dukungan institusional dalam adopsi solusi AI generatif di perguruan tinggi [11], [14], [19].

Temuan menunjukkan bahwa aspek teknis seperti rekayasa prompt dan pengelolaan konteks sisi-server berpotensi memengaruhi persepsi kredibilitas dan kenyamanan penggunaan. Kualitas desain sistem yang membantu menekan halusinasi dan menjaga kepatuhan struktur JSON pada format keluaran tampaknya turut berkontribusi terhadap terbentuknya persepsi positif pengguna, sebagaimana tercermin dalam indikator UTAUT dan AICA. Meskipun demikian, hubungan kausal antara aspek teknis dan persepsi pengguna masih memerlukan validasi lebih lanjut.

3.4. Keterbatasan hasil

Beberapa keterbatasan metodologis perlu dicatat. Pertama, jumlah indikator pada masing-masing konstruk masih terbatas pada tiga butir per konstruk, sehingga mengurangi fleksibilitas model pengukuran dan sensitivitas terhadap variabilitas respon. Kedua, ukuran sampel yang relatif moderat ($N = 321$) sudah memadai untuk analisis CFA dan estimasi model struktural SEM yang dilakukan, namun tetap terbatas untuk pengujian model yang lebih kompleks (misalnya multi-group SEM atau penambahan variabel mediasi dan moderasi). Ketiga, konteks penelitian yang berfokus pada satu institusi pendidikan di Batam membatasi generalisasi hasil ke lingkungan pendidikan tinggi lain dengan karakteristik berbeda. Keempat, seluruh pengukuran berbasis *self-report Likert* dapat mengandung potensi bias persepsi. Kelima, pengujian eksperimen A/B pada prototipe chatbot dilakukan dalam kondisi terkendali dengan jumlah skenario terbatas, sehingga performa aktual sistem di lingkungan produksi mungkin menunjukkan variasi tambahan. Terakhir, meskipun model struktural SEM telah diestimasi dan menunjukkan jalur pengaruh yang signifikan, desain penelitian yang bersifat potong-lintang (*cross-sectional*) dan berfokus pada satu institusi tetap membatasi kekuatan kesimpulan kausal. Hasil jalur struktural perlu dipahami sebagai dukungan empiris terhadap model teoritik, bukan sebagai bukti kausalitas yang final.

3.5. Implikasi ringkas

Secara praktis, pengembangan layanan informasi kampus berbasis *Generative AI* sebaiknya memprioritaskan orkestrasi teknis yang ketat pada sisi server guna meningkatkan keamanan dan keakuratan konten. Hasil eksperimen A/B menunjukkan bahwa prompt yang lebih terstruktur cenderung menghasilkan keluaran yang lebih akurat secara faktual serta sesuai format. Oleh karena itu, pendekatan ini patut dipertimbangkan untuk dipertahankan dalam pengembangan sistem. Di sisi lain, strategi implementasi perlu diiringi komunikasi institusional yang kuat guna membangun *endorsement sosial* dan kepercayaan terhadap kredibilitas sistem, mengingat kedua faktor tersebut terbukti cukup signifikan dalam memengaruhi niat adopsi pengguna [12], [13], [19]. Selain itu, dukungan infrastruktur digital dan kebijakan organisasi yang jelas menjadi

prasyarat penting keberlanjutan layanan AI di lingkungan pendidikan tinggi [11], [14], [19]. Temuan ini diharapkan dapat menjadi dasar bagi kampus dalam merancang tata kelola AI yang aman, inklusif, dan berorientasi pengguna.

4. KESIMPULAN

Penelitian ini mengembangkan dan mengevaluasi prototipe chatbot generatif untuk manajemen aktivitas mahasiswa dengan arsitektur *server-side context injection* serta mengkaji penerimaan pengguna menggunakan model UTAUT yang diperluas dengan konstruk AI Credibility Assurance (AICA). Dari sisi teknis, eksperimen A/B terhadap 100 kueri menunjukkan bahwa *engineered prompt* cukup signifikan ($p < 0.001$) untuk meningkatkan relevansi, kelengkapan, dan kegunaan keluaran dibanding *baseline prompt*, serta memperlihatkan tren akurasi faktual dan kepatuhan struktur JSON yang lebih tinggi. Reliabilitas penilaian dua rater yang tinggi mengonfirmasi bahwa rubrik evaluasi bersifat konsisten dan dapat diandalkan, pendekatan rekayasa prompt dan penyuntikan konteks sisi-server menunjukkan potensi dalam mengurangi halusinasi dan menjaga kualitas respons, sebagaimana tercermin dalam tren peningkatan akurasi faktual.

Dari sisi persepsi pengguna, survei terhadap 321 mahasiswa menunjukkan bahwa konstruk UTAUT yang diperluas (PE, EE, SI, FC, AICA, dan BI) memiliki reliabilitas internal dan validitas konstruk yang memadai. Measurement model enam faktor menunjukkan kecocokan yang baik serta mendukung validitas konvergen ($AVE \pm 0.66-0.77$; $CR \pm 0.85-0.91$) dan validitas diskriminan ($HTMT < 0.31$). Pada model struktural (DWLS), *Facilitating Conditions* ($\beta=0.228$; $p<0.001$) dan *Performance Expectancy* ($\beta=0.211$; $p<0.001$) merupakan prediktor terkuat terhadap *Behavioral Intention*, diikuti *Effort Expectancy* ($\beta=0.117$; $p<0.001$). *Social Influence* ($\beta=0.063$; $p=0.040$) dan *AI Credibility Assurance* ($\beta=0.070$; $p=0.033$) juga berasosiasi positif dan signifikan pada $\alpha=0.05$, namun dengan besaran efek relatif kecil.

Secara praktis, temuan penelitian ini mengindikasikan bahwa desain teknis dan pengelolaan konteks dapat berperan penting dalam membentuk persepsi kredibilitas dan kesiapan adopsi pengguna, khususnya dalam konteks kampus yang diteliti. Studi ini memberikan indikasi awal bahwa institusi pendidikan tinggi dapat mempertimbangkan integrasi praktik prompt engineering dan orkestrasi sisi-server dalam mengimplementasikan layanan AI, dengan tetap memperhatikan karakteristik lokal dan kesiapan infrastruktur masing-masing.

DAFTAR PUSTAKA

- [1]. Fatawi I, Asy'ari M, Hunaepi H, Samsuri T, Bilad MR. Empowering Language Models Through Advanced Prompt Engineering: A Comprehensive Bibliometric Review. *Indonesian Journal of Science & Technology*. 2024;9(2):441-462. doi:10.17509/ijost.v9i2.71481.
- [2]. Federiakin D, Molerov D, Zlatkin-Troitschanskaia O and Maur A. Prompt Engineering as a New 21st Century Skill. *Frontiers in Education*. 2024;9:1366434. doi:10.3389/feduc.2024.1366434.
- [3]. Chen B, Zhang Z, Langrené N, Zhu S. Unleashing the Potential of Prompt Engineering for Large Language Models. *Patterns*. 2025;6(5):101260. doi:10.1016/j.patter.2025.101260.
- [4]. Labadze L, Grigolia M, Machaidze L. Role of AI Chatbots in Education: Systematic Literature Review. *International Journal of Educational Technology in Higher Education*. 2023;20:56. doi:10.1186/s41239-023-00426-1.
- [5]. Bergman SA, et al. Leveraging Approximate Caching for Faster Retrieval-Augmented Generation. *Proceedings of the 5th Workshop on Machine Learning and Systems (EuroMLSys '25)*. Association for Computing Machinery; 2025. pp. 66-73. doi:10.1145/3721146.3721941.

- [6]. Priyadarshana YHPP, Senanayake A, Liang Z and Piumarta. Prompt Engineering for Digital Mental Health: A Short Review. *Frontiers in Digital Health*. 2024;6:1410947. doi:10.3389/fdgth.2024.1410947.
- [7]. Dermawan RD, Herdianto H. Meningkatkan Kinerja Output ChatGPT Melalui Teknik Prompt Engineering yang Dapat Dikustomisasi. *INNOVATIVE: Journal of Social Science Research*. 2024;4(1):10646-10664. doi:10.31004/innovative.v4i1.9067.
- [8]. Rizal MAS. Eksplorasi Penggunaan AI Generatif untuk Menciptakan Materi Pembelajaran Bahasa Indonesia yang Menarik dan Efektif. *INNOVATIVE: Journal of Social Science Research*. 2024;4(3):7080-7095. doi:10.31004/innovative.v4i3.10752.
- [9]. Erlina, et al. Penerapan Artificial Intelligence pada Aplikasi Chatbot sebagai Sistem Pelayanan dan Informasi Online pada Sekolah. *Journal of Information System and Technology (JOINT)*. 2023;4(3):421-430.
- [10]. Lau E, Yulianto A, Sabariman. Perancangan Chatbot pada PT Martindo Fine Foods Sebagai Solusi Peningkatan Efisiensi dan Layanan. *MADANI Jurnal Pengabdian Masyarakat dan Kewirausahaan*. 2024;2(4):167-174.
- [11]. Crompton H, Burke D. Artificial Intelligence in Higher Education: The State of the Field. *International Journal of Educational Technology in Higher Education*. 2023;20:22. doi:10.1186/s41239-023-00392-8.
- [12]. Marwah T, Agarwal PK, Gupta N. A Study on Acceptability of AI Assistant for Delivering Student Services in Private Universities of Delhi/NCR: A Study Based on the UTAUT Model. *European Economic Letters*. 2025;15(1): 3082-3103.
- [13]. Stöhr C, Ou AW, Malmström H. Perceptions and Usage of AI Chatbots Among Students in Higher Education Across Genders, Academic Levels and Fields of Study. *Computers & Education: Artificial Intelligence*. 2024;7:100259. doi:10.1016/j.caeai.2024.100259.
- [14]. Sugiono S. Proses Adopsi Teknologi Generative Artificial Intelligence dalam Dunia Pendidikan: Perspektif Teori Difusi Inovasi. *Jurnal Pendidikan dan Kebudayaan (JPKNK)*. 2024;9(1):110-133. doi:10.24832/jpnk.v9i1.4859.
- [15]. Ajiz MF, Ramadan MFS, Mutia HD, Yanuari PD. Pengembangan Aplikasi Chatbot Informasi Akademik Berbasis Web Menggunakan Metode Artificial Intelligence Markup Language (AIML). *Media Jurnal Informatika*. 2023;15(2):143-148. doi:10.35194/mji.v15i2.3316.
- [16]. Hron M, Obwegeser N. Why and How Is Scrum Being Adapted in Practice: A Systematic Review. *The Journal of Systems & Software*. 2022;183:111110. doi:10.1016/j.jss.2021.111110.
- [17]. Verwijns C, Russo D. A Theory of Scrum Team Effectiveness. *ACM Transactions on Software Engineering and Methodology*. 2023;32(3):Article 74. doi:10.1145/3571849.
- [18]. Alami A, Krancher O. How Scrum Adds Value to Achieving Software Quality? *Empirical Software Engineering*. 2022;27:165. doi:10.1007/s10664-022-10208-4.
- [19]. McGrath C, Farazouli A, Cerratto-Pargman T. Generative AI Chatbots in Higher Education: A Review of an Emerging Research Area. *Higher Education*. 2025;89:1533-1549. doi:10.1007/s10734-024-01288-w.
- [20]. Okonkwo CW, Ade-Ibijola A. Chatbots Applications in Education: A Systematic Review. *Computers and Education: Artificial Intelligence*. 2021;2:100033. doi:10.1016/j.caeai.2021.100033.
- [21]. Karabuğa B, et al. The Role of Artificial Intelligence (ChatGPT-4o) in Supporting Tumor Board Decisions. *Journal of Clinical Medicine*. 2025;14(10): 3535. doi:10.3390/jcm14103535.
- [22]. Marôco J. Factor Analysis of Ordinal Items: Old Questions, Modern Solutions? *Stats*. 2024;7:984-1001. doi:10.3390/stats7030060.
- [23]. Groskurth K, Bluemke M, Lechner CM. Why We Need to Abandon Fixed Cutoffs for Goodness-of-Fit Indices: An Extensive Simulation and Possible Solutions. *Behavior Research Methods*. 2024;56:3891-3914. doi:10.3758/s13428-023-02193-3.