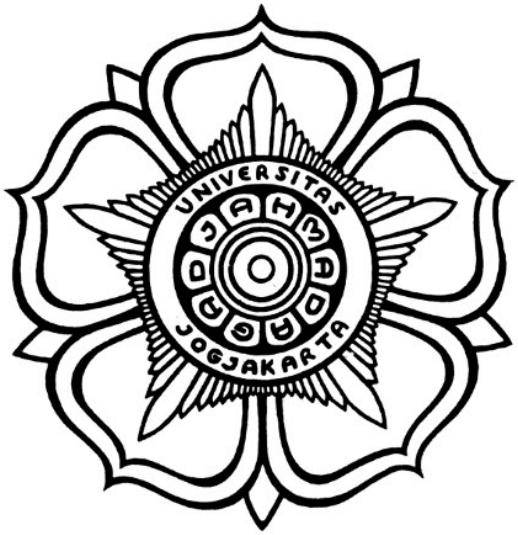
PENINGKATAN KUALITAS UMPAN BALIK PEDAGOGIS *LARGE LANGUAGE MODEL* BERBASIS *LEARNING ANALYTICS* PADA PLATFORM KANBAN UNTUK MENDUKUNG *SELF-REGULATED LEARNING* MAHASISWA

SKRIPSI



THE SUSTAINABLE DEVELOPMENT GOALS
Quality Education

Disusun oleh:

QORNAIN AJI
21/481767/TK/53170

PROGRAM SARJANA PROGRAM STUDI TEKNOLOGI INFORMASI
DEPARTEMEN TEKNIK ELEKTRO DAN TEKNOLOGI INFORMASI
FAKULTAS TEKNIK UNIVERSITAS GADJAH MADA
YOGYAKARTA
2025

HALAMAN PENGESAHAN

PENINGKATAN KUALITAS UMPAN BALIK PEDAGOGIS *LARGE LANGUAGE MODEL* BERBASIS *LEARNING ANALYTICS* PADA PLATFORM KANBAN UNTUK MENDUKUNG *SELF-REGULATED LEARNING* MAHASISWA

SKRIPSI

Diajukan Sebagai Salah Satu Syarat untuk Memperoleh
Gelar Sarjana Teknik
pada Departemen Teknik Elektro dan Teknologi Informasi
Fakultas Teknik
Universitas Gadjah Mada

Disusun oleh:

QORNAIN AJI
21/481767/TK/53170

Telah disetujui dan disahkan

Pada tanggal . . . . . .

Dosen Pembimbing I	Dosen Pembimbing II
Dr. Indriana Hidayah, S.T., M.T.	Syukron Abu Ishaq Alfarozi, S.T., Ph.D.
NIP 197905262002122001	**NIP 111199205202101102**

PERNYATAAN BEBAS PLAGIASI

Saya yang bertanda tangan di bawah ini :

Nama	: Qornain Aji
NIM	: 21/481767/TK/53170
Tahun terdaftar	: 2021
Program	: Sarjana
Program Studi	: Teknologi Informasi
Fakultas	: Teknik Universitas Gadjah Mada

Menyatakan bahwa dalam dokumen ilmiah Skripsi ini tidak terdapat bagian dari karya ilmiah lain yang telah diajukan untuk memperoleh gelar akademik di suatu lembaga Pendidikan Tinggi, dan juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang/lembaga lain, kecuali yang secara tertulis disitasi dalam dokumen ini dan disebutkan sumbernya secara lengkap dalam daftar pustaka.

Dengan demikian saya menyatakan bahwa dokumen ilmiah ini bebas dari unsur-unsur plagiasi dan apabila dokumen ilmiah Skripsi ini di kemudian hari terbukti merupakan plagiasi dari hasil karya penulis lain dan/atau dengan sengaja mengajukan karya atau pendapat yang merupakan hasil karya penulis lain, maka penulis bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Yogyakarta, 20 November 2025

Qornain Aji
21/481767/TK/53170

HALAMAN PERSEMBAHAN

Tugas akhir ini kupersembahkan kepada Papah Mamah & Kakak (serta Mochi) dengan penuh cinta dan rasa terima kasih!

Karya ini merupakan bukti perjuangan dan doa kalian selama ini yang tiada henti sebagai mahasiswa Teknologi Informasi angkatan 2021. Terima kasih atas segala dukungan, motivasi, dan kasih sayang yang selalu kalian berikan kepadaku. Semoga Allah membalas semua kebaikan kalian dengan balasan yang berlipat ganda. Aamiin.

KATA PENGANTAR

Puji syukur kehadirat Allah SWT atas limpahan rahmat, karunia, serta petunjuk-Nya sehingga tugas akhir berupa penyusunan skripsi ini telah terselesaikan dengan baik dan tepat waktu. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana di Program Studi Teknologi Informasi.

Dalam proses penyusunan tugas akhir yang penuh tantangan ini, penulis telah banyak mendapatkan arahan, bantuan, dukungan moral, mental, dan juga ilmu pengetahuan dari berbagai pihak. Oleh karena itu, pada kesempatan ini dengan segala kerendahan hati, penulis mengucapkan terima kasih yang sebesar-besarnya kepada

1. Bapak Prof. Ir. Hanung Adi Nugroho, S.T., M.E., Ph.D., IPM., SMIEEE., selaku Ketua Departemen Teknik Elektro dan Teknologi Informasi FT UGM, atas dukungan dan arahan yang tak ternilai selama masa studi.
2. Bapak Ir. Lesnanto Multa Putranto, S.T., M.Eng., Ph.D., IPM., SMIEEE., selaku Sekretaris Departemen, terima kasih atas berbagai kemudahan administrasi dan fasilitas yang telah diberikan selama penulis menempuh studi.
3. Dr. Indriana Hidayah, S.T., M.T., sebagai Dosen Pembimbing Utama, yang telah meluangkan waktu, mencurahkan tenaga dan pikiran, memberikan bimbingan intensif, arahan yang tajam, serta motivasi berkelanjutan sejak awal hingga akhir penyelesaian skripsi ini.
4. Syukron Abu Ishaq Alfarozi, S.T., Ph.D., sebagai Dosen Pembimbing Pendamping, yang telah memberikan arahan berharga, masukan konstruktif, serta dukungan moral selama proses penyusunan skripsi ini.
5. Ayah, Teguh Bharata Adji, S.T., M.T., M.Eng., Ph.D., yang selalu mengingatkan untuk salat lima waktu tepat waktu, serta berkorban mencari nafkah demi pendidikan anak-anaknya, mendidik menjadi pribadi yang kuat & tegar. Terima kasih atas doa, dukungan moral, motivasi, dan kasih sayang yang tiada henti selama ini.
6. Ibu, apt. RR. Sabtanti Harimurti, S.Si., M.Sc., Ph.D., yang selalu memberikan kasih sayang, memberikan motivasi, serta doa yang tak pernah putus untuk kesuksesan anak-anaknya. Terima kasih atas segala pengorbanan Ibu selama ini.
7. Kakak tercinta, Lutfi Aji S.T., yang selalu menjadi panutan, mengenalkan arti tentang hidup, mengajar dan mengajak bermain bola basket, serta memberikan dukungan moral dan motivasi selama ini.
8. Teman-teman "Darurat Bang" TIF 2021, "GMM", "Anak Bunda", terima kasih atas kebersamaan, dukungan, motivasi, serta canda tawa yang telah kita bagi bersama selama ini dari bangku SMA hingga perkuliahan.
9. Organisasi KMTETI, BEM KM FT Andal, Tim Basket DTETI, DTETI 21, WebDev TC 2023, dan teman-teman ITS*Research Group* atas pengalaman berharga, kebersamaan, suka dan duka, serta ilmu yang telah kita bagi bersama selama ini.

Akhir kata, penulis berharap semoga skripsi ini dapat memberikan manfaat yang sebesar-besarnya bagi perkembangan ilmu pengetahuan, khususnya di bidang Teknologi Informasi, dan dapat memberikan inspirasi serta motivasi bagi kita semua, *aamiin*.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR SINGKATAN

[SAMPLE]

b	=	bias
$K(x_i, x_j)$	=	fungsi kernel
y	=	kelas keluaran
C	=	parameter untuk mengendalaikan besarnya pertukaran antara penalti variabel slack dengan ukuran margin
L_D	=	persamaan Lagrange dual
L_P	=	persamaan Lagrange primal
w	=	vektor bobot
x	=	vektor masukan
ANFIS	=	Adaptive Network Fuzzy Inference System
ANSI	=	American National Standards Institute
DAG	=	Directed Acyclic Graph
DDAG	=	Decision Directed Acyclic Graph
HIS	=	Hue Saturation Intensity
QP	=	Quadratic Programming
RBF	=	Radial Basis Function
RGB	=	Red Green Blue
SV	=	Support Vector
SVM	=	Support Vector Machines

INTISARI

Intisari ditulis menggunakan bahasa Indonesia dengan jarak antar baris 1 spasi dan maksimal 1 halaman. Intisari sekurang-kurangnya berisi tentang latar belakang dan tujuan penelitian, metodologi yang digunakan, hasil penelitian, kesimpulan dan implikasi, dan Kata kunci yang berhubungan dengan penelitian.

Kata Kunci ditulis maksimal 5 kata yang paling berhubungan dengan isi skripsi. Silakan mengacu pada ACM / IEEE *Computing classification* jika Anda adalah mahasiswa Sarjana TI http://www.acm.org/about/class/ atau mengacu kepada IEEE keywords http://www.ieee.org/documents/taxonomy_v101.pdf jika Anda berasal dari Prodi Sarjana TE.

Kata kunci : Kata kunci 1, Kata kunci 2, Kata kunci 3, Kata kunci 4, Kata kunci 5

Contoh Abstrak Teknik Elektro:

"Penelitian ini bertujuan untuk mengembangkan sistem pengendalian suhu ruangan dengan menggunakan microcontroller. Metodologi yang digunakan adalah desain sirkuit, implementasi sistem pengendalian, dan pengujian performa. Hasil penelitian menunjukkan bahwa sistem pengendalian suhu ruangan yang dikembangkan mampu mengendalikan suhu ruangan dengan akurasi sebesar ±0,5°C. Kesimpulan dari penelitian ini adalah sistem pengendalian suhu ruangan yang dikembangkan efektif dan efisien.

Kata kunci: microcontroller, sistem pengendalian suhu, akurasi."

Contoh Abstrak Teknik Biomedis:

"Penelitian ini bertujuan untuk mengevaluasi keefektifan prototipe alat pemantau denyut jantung berbasis elektrokardiogram (ECG) untuk pasien jantung. Metodologi yang digunakan meliputi desain dan pembuatan prototipe, pengujian dengan pasien, dan analisis data. Hasil penelitian menunjukkan bahwa prototipe alat pemantau denyut jantung berbasis ECG memiliki akurasi yang baik dan mampu memantau denyut jantung pasien secara efektif. Kesimpulan dari penelitian ini adalah prototipe alat pemantau denyut jantung berbasis ECG merupakan solusi yang efektif dan efisien untuk memantau pasien jantung.

Kata kunci: elektrokardiogram, alat pemantau denyut jantung, akurasi."

Contoh Abstrak Teknologi Informasi:

"Penelitian ini bertujuan untuk mengevaluasi keamanan dan privasi pengguna aplikasi media sosial terpopuler. Metodologi yang digunakan meliputi analisis kebijakan privasi dan pengaturan keamanan, pengujian penetrasi, dan survei pengguna. Hasil penelitian menunjukkan bahwa beberapa aplikasi media sosial memiliki kebijakan privasi yang kurang jelas dan rendahnya tingkat keamanan. Kesimpulan dari penelitian ini adalah pentingnya meningkatkan kebijakan privasi dan tingkat keamanan pada aplikasi media sosial untuk melindungi privasi dan data pengguna.

Kata kunci: media sosial, keamanan, privasi, pengguna."

ABSTRACT

Abstract ditulis italic (miring) menggunakan bahasa Inggris dengan jarak antar baris 1 spasi dan maksimal 1 halaman. Abstract adalah versi Bahasa Inggris dari intisari. Abstract dapat ditulis dalam beberapa paragraf. Baris pertama paragraph harus menjorok ke dalam sekitar 1 cm. Tidak dsarankan menggunakan mesin penerjemah melainkan tulis ulang.

Keywords : Keyword 1, Keyword 2, Keyword 3, Keyword 4, Keyword 5

BAB I
PENDAHULUAN

1.1 Latar Belakang

Dunia pendidikan tinggi saat ini menuntut mahasiswa untuk bisa belajar secara mandiri. Namun banyak dari mereka belum siap menghadapi tuntutan tersebut. Hal ini menimbulkan kesenjangan yang serius dalam keterampilan *Self-Regulated Learning* (SRL), yang dapat menghambat keberhasilan belajar dan mengganggu pemerataan akses pendidikan [3, 4]. Tantangan ini justru diperburuk oleh teknologi digital yang menjadi ciri khas pembelajaran masa kini. Meskipun menawarkan fleksibilitas, teknologi tersebut juga menghadirkan distraksi yang signifikan dan menuntut disiplin diri yang lebih besar [5]. Kegagalan institusi dalam memberikan dukungan yang memadai untuk pengembangan SRL bukan hanya masalah akademik, melainkan juga masalah sosial yang berdampak pada prospek karier jangka panjang mahasiswa dan efektivitas sistem pendidikan [3].

Untuk mengatasi kesenjangan SRL, Zimmerman mengemukakan kerangka kerja yang mencakup tahap perencanaan (*forethought*), pelaksanaan (*performance*), dan refleksi diri (*self-reflection*) di mana SRL merupakan gabungan strategi kognitif, motivasional, dan perilaku [3, 6]. Individu dengan kemampuan regulasi diri yang efektif terbentuk melalui penetapan tujuan, perancangan strategi, pemantauan diri secara aktif, dan evaluasi diri yang bijaksana. Sejumlah besar penelitian mengonfirmasi hubungan kuat antara keterampilan ini dan pencapaian akademik [3, 4]. Namun, pengembangannya kerap terhambat oleh faktor personal, kontekstual, dan sosial, mulai dari manajemen waktu yang buruk dan distraksi digital hingga rendahnya kesadaran metakognitif. Metakognisi sendiri merupakan pengetahuan, kesadaran, dan regulasi atas pemikiran diri sendiri, yang berperan penting dalam membangun kesadaran mahasiswa akan proses belajarnya [7]. Hambatan tersebut tersebut menyebabkan intervensi terstruktur diperlukan untuk membimbing mahasiswa menjalani proses yang kompleks ini [5, 8].

Metode tradisional yang paling efektif untuk mengembangkan keterampilan SRL adalah umpan balik personal dari ahli manusia. Umpan balik berkualitas tinggi yang berfokus tidak hanya pada jumlah tugas yang diselesaikan, tetapi juga pada proses belajar dan regulasi diri mahasiswa merupakan katalis kuat untuk perbaikan [15, 16] [9, 10]. Namun "standar emas" ini menghadapi krisis skalabilitas dengan meningkatnya jumlah mahasiswa dan kebutuhan dukungan yang beragam membuat model bimbingan secara personal sulit diterapkan dalam skala besar. Hal ini diperkuat dengan studi-studi yang dilakukan dan menunjukkan bahwa tekanan anggaran mendorong ukuran kelas yang semakin besar, sehingga interaksi personal antara dosen dan mahasiswa sangat terbatas [11, 12].

Ini menciptakan kebutuhan mendesak akan solusi inovatif berbasis teknologi yang dapat melengkapi keahlian manusia.

Salah satu teknologi tersebut adalah penggunaan *Large Language Models* (LLM). LLM dapat digunakan sebagai tutor virtual yang mampu memberi umpan balik yang cepat dan konsisten. LLM telah digunakan di berbagai penelitian dalam konteks pendidikan dengan bermacam terobosan. Temuan terbaru memperlihatkan bahwa ketika LLM digunakan untuk penilaian dan umpan balik, *artificial intelligence* (AI) cenderung murah hati dalam memberikan nilai sehingga nilainya lebih tinggi daripada nilai yang diberikan ahli manusia. Sementara itu penilaian antar rekan dan dosen lebih rendah dan sesuai dengan performa mahasiswa. Umpan balik AI sering terstruktur, tetapi tetap memerlukan pengawasan ahli manusia agar relevan secara pedagogis sehingga pendekatan hibrida (penggabungan AI dan ahli manusia) direkomendasikan [13]. Pendekatan hibrida direkomendasikan karena dapat menggabungkan kecepatan dan skalabilitas AI dengan penguasaan konteks, pemahaman empati, dan nuansa pedagogis dari penilaian manusia [13]. Rekomendasi tersebut menegaskan bahwa respons LLM masih membutuhkan pengawasan pakar agar relevan. Di sisi lain, upaya standarisasi evaluasi tutor-AI menunjukkan bahwa meskipun model terkini seperti GPT-4 kuat dalam menjawab pertanyaan, mereka cenderung terlalu cepat memberikan jawaban kepada mahasiswa dan kurang menuntun proses. Fakta ini menunjukkan bahwa LLM belum ideal sebagai tutor dan belum dapat menggantikan ahli manusia dalam penilaian kualitas pedagogis [14].

Oleh karena itu, skripsi ini berfokus pada peningkatan kualitas performa respons LLM sebagai pemberi umpan balik agar mendekati mutu ahli manusia secara kuantitatif dan kualitatif. Peningkatan dilakukan melalui kerangka evaluasi yang menggabungkan acuan ahli manusia sebagai *ground truth* dan penilaian kualitas yang menilai dimensi-dimensi pedagogis relevan [13, 14]. Untuk mendukung proses pengukuran tersebut, skripsi ini memanfaatkan jejak proses belajar yang terstruktur. Data pembelajaran mahasiswa yang dipresentasikan ke dalam bentuk papan Kanban dipakai sebatas alat operasional untuk menata dan mengekstraksi data proses (perpindahan kartu, checklist, waktu pengerjaan) yang dapat dipetakan ke fase SRL (perencanaan-*monitoring*-refleksi) [15]. Penggunaan papan Kanban hanya sebagai salah satu instrumen pengaya data agar evaluasi dan perbaikan kualitas umpan balik LLM bisa berjalan lebih terarah. Pada penelitian ini, dijelaskan pada bab berikutnya bahwa fokus kebaruan terletak pada rangkaian evaluasi dan perbaikan *prompt* LLM. Penelitian ini akan memperbaiki tiga fungsi pedagogis utama yakni feedback, dukungan motivasional, dan dukungan apresiatif dengan tujuan praktis yaitu menyamai kualitas ahli manusia pada metrik kuantitatif sekaligus memenuhi standar kualitas kualitatif [9, 10, 13, 14]. Dengan membandingkan keluaran LLM dan pakar ahli, studi ini menilai peningkatan performa LLM agar dapat menjadi pertimbangan dalam penggunaan sebagai agen pedagogis pelengkap mahasiswa yang masif dan

personal.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, penelitian ini mengidentifikasi Secara spesifik, penelitian ini berfokus pada pengembangan umpan balik yang menargetkan tiga fungsi pedagogis yaitu *feedback*, *motivation*, dan *appreciation*. Rumusan masalah yang diangkat dalam penelitian ini adalah sebagai berikut:

1. Belum layaknya kualitas umpan balik LLM terhadap proses SRL mahasiswa yang direpresentasikan melalui data papan Kanban jika dibandingkan dengan umpan balik ahli manusia.
2. Butuhnya metode evaluasi yang melibatkan acuan ahli manusia sebagai *ground truth* dan penilaian kualitas yang menilai dimensi-dimensi pedagogis relevan untuk mengukur kualitas umpan balik LLM terhadap proses SRL mahasiswa.
3. Belum jelas pengaruh ukuran model LLM terhadap kualitas umpan balik yang dihasilkan di mana semakin besar ukuran model LLM, semakin mahal biayanya.

1.3 Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Merancang metode yang sesuai untuk meningkatkan performa LLM dalam memberikan umpan balik terhadap proses SRL mahasiswa jika dibandingkan dengan ahli manusia.
2. Mengevaluasi kualitas umpan balik yang dihasilkan LLM dari berbagai skema metode yang dirancang terhadap proses SRL mahasiswa baik secara kuantitatif maupun kualitatif dengan umpan balik ahli manusia.
3. Melihat pengaruh besar ukuran model LLM terhadap kualitas umpan balik yang dihasilkan.

1.4 Batasan Penelitian

Penelitian ini berfokus untuk melihat peningkatan performa dari LLM ketika diberikan perlakuan berupa metode *context engineering* yang memanfaatkan data *learning analytics*. Upaya untuk menyederhanakan proses ini dilakukan dengan cara memberikan batasan penelitian yang jelas dan ditetapkan oleh peneliti dengan mempertimbangkan aspek-aspek yang relevan. Berikut adalah batasan-batasan tersebut:

1.4.1 Objek Penelitian

(a) Ruang lingkup mencakup analisis kemampuan model bahasa besar pada tugas berbahasa Indonesia dengan konteks pembelajaran berbasis papan Kanban.

(b) Data pembelajaran yang digunakan berupa data sintetis yang disusun untuk memicu respons dari model dan dari pakar manusia. Pemakaian data sintetis dinilai m emadai karena keterbatasan waktu pengambilan data ril serta karena fokus utama penelitian adalah perbandingan kualitas respons model dan respons pakar.

(c) Model yang diteliti merupakan model yang tersedia secara *open source* dan diakses melalui Groq API. Dua model yang digunakan adalah Llama 3.1 8B Instant 128k dan Llama 3.3 70B Versatile 128k.

1.4.2 Metode Penelitian

Penelitian menggunakan rancangan eksperimen komparatif tanpa pelatihan ulang pada model. Seluruh prosedur mengandalkan rekayasa konteks yang mencakup perancangan templat *prompt* penetapan instruksi dan penyediaan konteks dari data Kanban berbahasa Indonesia. Langkah penelitian meliputi penyusunan skenario soal dan konteks berbasis Kanban pembuatan respons oleh kedua model melalui Groq API pengumpulan respons pakar manusia pada skenario yang sama serta evaluasi kuantitatif dan penilaian kualitatif oleh pakar. Reproduksibilitas dijaga melalui pencatatan parameter *decoding* penetapan *seed* serta pelaporan lingkungan komputasi.

1.4.3 Waktu dan Tempat Penelitian

Penelitian dilaksanakan pada Juni 2025 hingga Oktober 2025. Proses perancangan dan analisis dilakukan dengan memanfaatkan layanan Groq untuk inferensi model.

1.4.4 Populasi dan Sampel

Populasi penelitian mencakup seluruh respons atas skenario berbahasa Indonesia yang relevan dengan konteks pembelajaran berbasis Kanban. Sampel terdiri atas sekumpulan skenario tugas yang dipilih secara purposif beserta pasangan respons model dan respons pakar manusia. Jumlah *rater* pakar jumlah skenario serta pembagian data ditetapkan agar memadai untuk analisis kuantitatif dan penilaian kualitatif.

1.4.5 Variabel

(a) **Variabel bebas** mencakup jenis model Llama 3.1 8B Instant 128k dan Llama 3.3 70B Versatile 128k serta strategi rekayasa konteks yang digunakan pada *prompt* dan penyajian *learning analytics* berbasis Kanban .

(b) **Variabel terikat** adalah kualitas keluaran yang diukur dengan metrik kuantitatif seperti *accuracy precision recall macro F1* serta ukuran kesesuaian teks yang relevan.

(c) **Variabel kendali** meliputi penggunaan bahasa Indonesia pada seluruh data karakteristik skenario Kanban, parameter seperti *temperature* dan *top-p random seed* serta prosedur evaluasi yang seragam.

1.4.6 Keterbatasan Penelitian

1. Data pembelajaran berbasis papan Kanban yang digunakan merupakan data sintetis yang disusun untuk keperluan pengambilan data dari pakar manusia dan model. Data sintetis dipilih karena keterbatasan waktu pengambilan data ril serta karena fokus penelitian berada pada perbandingan respons.
2. Penelitian hanya menggunakan model yang telah tersedia dan bersifat *open source* guna efisiensi waktu dan biaya.
3. Model diakses melalui Groq API sebagai sarana inferensi.
4. Model yang digunakan terbatas pada dua model yaitu Llama 3.1 8B Instant 128k dan Llama 3.3 70B Versatile 128k.
5. Tidak dilakukan pelatihan ulang pada model sehingga seluruh prosedur bersifat rekayasa konteks.
6. Seluruh kegiatan dilakukan pada Juni 2025 hingga Oktober 2025.
7. Analisis difokuskan pada kemampuan berbahasa Indonesia sehingga seluruh data yang digunakan berbahasa Indonesia.
8. Sumber data utama berasal dari keluaran model dan respons pakar manusia yang dianalisis sedangkan data pembelajaran Kanban berfungsi sebagai pendukung penyusunan konteks.
9. Luaran penelitian berfokus pada identifikasi metode optimisasi yang paling baik berdasarkan pengukuran kuantitatif yang selanjutnya diuji kepada pakar manusia untuk penilaian kualitatif.

1.5 Manfaat Penelitian

Penelitian ini menawarkan wawasan bagaimana penerapan LLM sebagai pendamping dalam proses SRL bagi mahasiswa bisa menjadi alternatif yang efektif dalam meningkatkan kualitas pembelajaran. Selain itu, penelitian ini juga dapat menjadi dasar untuk penelitian lebih lanjut untuk peningkatan performa respons LLM terutama pada bidang pedagogis sehingga seiring berjalannya waktu dapat memberikan hasil yang lebih optimal.

1.6 Sistematika Penulisan

Sistematika penulisan berisi pembahasan apa yang akan ditulis di setiap bab. Sistematika pada umumnya berupa paragraf yang setiap paragraf mencerminkan bahasan setiap Bab. Contoh:

Bab I membahas tentang pendahuluan yang berisi latar belakang, perumusan masalah dan tujuan penelitian.

Bab II berisi tentang metodologi penelitian yang terdiri dari desain penelitian, sumber data, Teknik pengumpulan data dan Teknik analisis data.

Dan seterusnya.

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

Bab ini menyajikan tinjauan komprehensif yang disusun secara sistematis untuk membangun landasan penelitian. Tinjauan dimulai dengan analisis terhadap penelitian-penelitian terdahulu, dilanjutkan dengan komparasi kritikal, dan diakhiri dengan pembahasan dasar teori yang relevan.

Pertama, dibahas terlebih dahulu kumpulan penelitian sebelumnya yang berfokus pada upaya meningkatkan kualitas umpan balik dari Large Language Models (LLMs) dalam konteks pendidikan. Termasuk di dalamnya adalah berbagai strategi optimasi yang telah diusulkan serta metode-metode evaluasi yang digunakan untuk mengukur efektivitas umpan balik yang dihasilkan. Kedua, dilakukan komparasi mendalam terhadap penelitian-penelitian tersebut untuk mengidentifikasi kesenjangan yang ada di dalam literatur. Analisis komparatif ini membantu memetakan posisi penelitian ini dalam lanskap penelitian yang lebih luas. Ketiga, dibahas metode-metode evaluasi yang telah digunakan dalam studi-studi sebelumnya. Hal ini mencakup evaluasi baik dari segi kuantitatif maupun kualitatif. Keempat, berdasarkan identifikasi kesenjangan tersebut, disajikan dasar teori yang melandasi penelitian ini. Pembahasan mencakup peran LLM dalam dunia pendidikan, kerangka teoritis Self-Regulated Learning (SRL), konsep Prompt Engineering dan Context Engineering, pendekatan Learning Analytics beserta metode-metode analitik yang relevan untuk konteks pembelajaran, serta metode evaluasi yang dapat digunakan di dalam penelitian ini.

Melalui struktur ini, bab ini tidak hanya memberikan fondasi teoretis yang kuat, tetapi juga memperjelas kontribusi penelitian dalam mengisi celah yang teridentifikasi dari penelitian-penelitian sebelumnya.

2.1 Tinjauan Pustaka

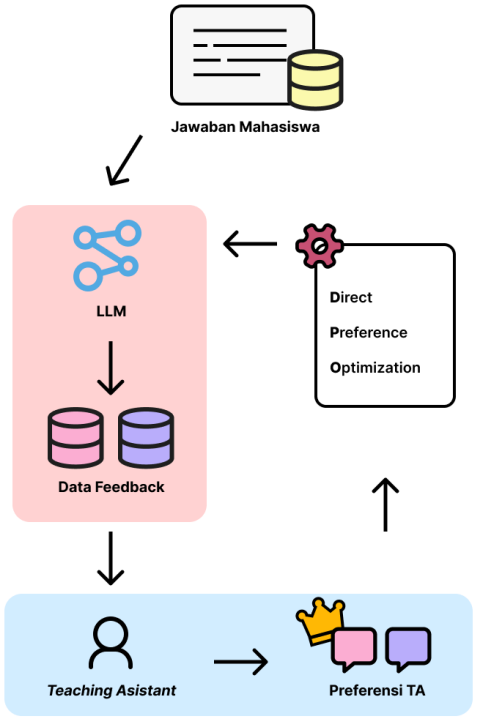
Bagian ini menyajikan tinjauan terhadap penelitian-penelitian sebelumnya yang menjadi landasan bagi pengembangan sistem umpan balik LLM dalam pendidikan. Tinjauan ini mencakup dua fokus utama yaitu strategi peningkatan kualitas umpan balik LLM dan metode evaluasi yang digunakan.

Melalui analisis komparatif terhadap berbagai pendekatan yang ada, diidentifikasi pola, tren, dan kesenjangan dalam literatur. Pemetaan ini tidak hanya memberikan konteks bagi penelitian ini, tetapi juga memperjelas kontribusi unik yang ditawarkan dalam mengatasi keterbatasan-keterbatasan yang masih ada pada penelitian-penelitian sebelumnya.

2.1.1 Peningkatan Umpan Balik LLM dengan *Direct Preference Optimization* (DPO)

Juliette Woodrow, Sanmi Koyejo, dan Chris Piech (Stanford), mengajukan makalah "*Improving Generative AI Student Feedback: Direct Preference Optimization with Teachers in the Loop*" pada EDM 2025 [1]. Mereka memotret persoalan klasik pada umpan balik berbasis LLM yaitu, walau model mampu menghasilkan teks yang koheren, gaya, ketelitian, dan terminologi, sering kali tidak selaras dengan preferensi pengajar dan konteks spesifik mata kuliah. "*Feedback Alignment Challenge*" disebutkan sebagai tantangan menyesuaikan keluaran AI dengan ketelitian, terminologi, dan gaya respons yang diharapkan tim pengajar, sehingga *feedback* terasa "*course-specific*", bukan *feedback* generik.

Solusi yang ditawarkan adalah *Direct Preference Optimization* (DPO) dengan guru atau *teaching assistant* (TA) di dalam loopnya. Inti dari DPO adalah selama proses penilaian, sistem menampilkan dua kandidat umpan balik AI untuk setiap jawaban mahasiswa, asisten dosen dapat memilih, mengedit, atau menulis ulang. Pilihan ini menjadi pasangan preferensi (menangkalah) yang kemudian dipakai untuk *fine-tuning* model menggunakan DPO, tanpa perlu membangun reward model seperti pada *Reinforcement Learning from Human Feedback* (RLHF) pada Subbab 2.1.3 sehingga alurnya lebih sederhana dan langsung menekan probabilitas keluaran yang disukai dibanding yang tidak disukai. Dirancang pipa tiga tahap yaitu koleksi preferensi saat grading, pelatihan DPO antar tugas, dan inferensi untuk tugas berikutnya sehingga mewujudkan sistem *self-improving* yang semakin selaras dengan preferensi pengajar dari satu *assignment* ke *assignment* berikutnya. Ilustrasi alur DPO dapat dilihat pada Gambar 2.1.



Gambar 2.1. Alur Kerja *Direct Preference Optimization* (DPO) dengan Guru di Dalam Loop [1]

Dalam implementasinya, mereka melakukan *fine-tuning* pada model Llama 3.1 8B Instruct menggunakan HuggingFace DPOTrainer. Proses pelatihan berlangsung sekitar tujuh jam pada tiga GPU A6000 dengan rata-rata 1.408 pasangan preferensi per gelombang. Pada tahap noticing, yaitu pengumpulan observasi atas pekerjaan mahasiswa, mereka memanfaatkan rubrik, tes otomatis, dan juga bantuan model LLM lain (GPT-4o). Hasil dari tahap ini kemudian disusun ke dalam prompt terstruktur agar model dapat menghasilkan respons yang lebih tepat sasaran.

Pengujian pendekatan ini dilakukan pada dua skenario. Pertama, studi terkontrol buta yakni digunakan 10 evaluator ahli (staf pengajar) membandingkan umpan balik dari beberapa model pada kiriman mahasiswa yang dianonimkan di mana mereka tidak mengetahui identitas model dan menilai preferensi, *Insightfulness*, serta *Naturalness*. Kedua, deployment nyata di dua penyelenggaraan mata kuliah universitas besar (≈ 330 dan ≈ 289 mahasiswa; >10 *teaching assistants* (TA) per semester), di mana setiap generasi model dilatih dari preferensi tugas sebelumnya lalu dipakai pada tugas berikutnya. Dilakukan juga pencatatan potensi *confounding variable* dalam deployment (mis. setiap soal dinilai oleh satu TA sehingga preferensi pribadi bisa memengaruhi pemilihan *feedback*; variasi soal antar generasi menyulitkan isolasi efek model).

Untuk evaluasi, dilakukan pemadanan penilaian manusia secara buta dan penilai-

an otomatis berbasis *critic model*. Pada sisi otomatis, mereka mengadopsi kerangka Scarlatos et al. [16] guna menilai lima dimensi yaitu, *Correctness, Not-revealing-the-answer, Suggestions, Positive tone*, serta skor total. Kerangka ini juga harusnya memuat *"diagnosing errors"* namun tidak difokuskan di hasil tabel karena tidak sejalan dengan tugas utama studi ini. Di luar itu, dilakukan juga pengenalan *critic model* baru dengan tiga dimensi yaitu *Assertiveness*, *Accuracy*, dan *Helpfulness* yang lebih umum untuk berbagai tipe jawaban (benar/salah). Kedua kerangka ini menggunakan GPT-4o sebagai evaluator otomatis; prompts dan kode penilaian dibuka untuk umum.

Hasil studi terkontrol menunjukkan bahwa model DPO lebih disukai dibanding baseline kuat GPT-4o dengan persentase 56,8% berbanding 40,2% sementara 3% responden memilih netral. Pada metrik *Insightfulness*, DPO juga memperoleh skor lebih tinggi secara signifikan dibanding GPT-4o (rata-rata 3,13 berbanding 2,79) maupun model SFT (2,82). Sebaliknya, pada metrik *Naturalness* tidak ditemukan perbedaan signifikan antar model. Ulasan kualitatif evaluator menegaskan bahwa alasan memilih DPO adalah karena tingkat kekhususan yang lebih tinggi, relevansi terhadap masalah, nada yang mendorong, serta saran yang lebih dapat ditindaklanjuti. Sementara itu, alasan utama dipilihnya GPT-4o adalah kejelasan dan keringkasan, sementara DPO terkadang dinilai terlalu panjang atau agak canggung dari segi gaya.

Pada evaluasi otomatis (kerangka Scarlatos), DPO melampaui GPT-4o pada 4/5 metrik *Correctness* (0,932 ± 0,004 vs 0,862 ± 0,006), *Not-revealing-the-answer* (0,980 ± 0,002 vs 0,973 ± 0,003), *Positive tone* (0,989 ± 0,002 vs 0,905 ± 0,005), dan skor total (0,724 ± 0,004 vs 0,664 ± 0,005). Satu-satunya area yang lebih baik pada GPT-4o adalah *Suggestions* (GPT-4o 0,229 ± 0,007 vs DPO 0,163 ± 0,006), menandakan bahwa keluaran GPT-4o relatif lebih banyak memberi saran eksplisit, sementara DPO unggul dalam ketepatan, menjaga agar tidak membocorkan jawaban, dan mempertahankan nada positif. Di critic model baru, sebaran titik menunjukkan klaster DPO lebih konsisten di area akurasi dan helpfulness yang tinggi, sedangkan GPT-4o menampilkan variasi lebih besar berikut klaster pada wilayah rendah akurasi dan rendah helpfulness; secara kuantitatif, umpan balik GPT-4o 8,73 kali lebih mungkin tergolong unhelpful dan 7 kali lebih mungkin inaccurate dibanding DPO (tingkat assertiveness relatif sebanding). Hasil deployment memperlihatkan kelayakan integrasi di kelas besar dan penguatan preferensi TA terhadap DPO dalam seting riil, seraya membuka jalan untuk monitoring otomatis dan eksplorasi awal fairness (gender) yang dapat diperluas ke demografi lain.

2.1.2 Peningkatan *Feedback* LLM dengan *Prompt Engineering*

Artikel ini ditulis oleh Lucas Jasper Jacobsen dan Kira Elena Weber (Universität Hamburg) [17]. Tujuan utamanya ada dua: (1) mengidentifikasi jenis prompt seperti apa yang dibutuhkan agar LLM dapat menghasilkan umpan balik berkualitas tinggi di

pendidikan guru, dan (2) membandingkan kualitas umpan balik LLM dengan umpan balik manusia dari kelompok novice (calon guru/preservice teachers) dan expert. Latar masalah yang mereka tekankan: studi empiris yang langsung membandingkan kualitas umpan balik LLM vs manusia masih jarang, dan belum ada manual/kerangka berbasis teori untuk menilai kualitas prompt di konteks pendidikan tinggi, padahal umpan balik berkualitas penting namun sumber daya manusia dan waktu dosen sering terbatas.

Sebagai respons, dilakukan pengembangan sebuah manual prompt berbasis teori. Manual ini merangkum kategori-kategori kualitas prompt (mencakup aspek konteks/peran-audiens-medium, misi, kejernihan dan spesifikasi—misalnya format, kekonkretan, spesifik domain, dan logika instruksi). Berdasarkan manual tersebut, mereka merancang tiga versi prompt dengan kualitas rendah-menengah-tinggi. Seluruh generasi umpan balik LLM menggunakan ChatGPT-4, dengan alasan kesesuaian praktik nyata di pendidikan tinggi saat penelitian dilakukan. Gagasannya sederhana: jika mutu prompt dapat disistematiskan, kualitas umpan balik LLM akan lebih konsisten dan bisa dibandingkan secara adil dengan umpan balik manusia.

Tugas yang dipakai berupa learning goal di topik geometri (mengenali segitiga siku-siku dan teorema Pythagoras), lengkap dengan tiga jenis kesalahan yang umum pada perumusan tujuan pembelajaran. Dari tiga kualitas prompt (rendah-menengah-tinggi), dihasilkan 20 umpan balik per prompt (total 60 keluaran LLM) yang kemudian dikodekan oleh tiga coder terlatih. Untuk pembandingan antar penyedia umpan balik, kelompok novice dan expert juga diminta menulis umpan balik dengan menggunakan prompt terbaik (kualitas tinggi), sehingga perbandingan dengan LLM berlangsung pada kondisi instruksi yang optimal. Desain ini memungkinkan isolasi efek "kualitas prompt" pada kinerja LLM sekaligus membangun baseline manusia yang kuat pada kondisi yang sama.

Kualitas umpan balik dianalisis dengan skema kodifikasi yang mencakup dimensi-dimensi pedagogis seperti concreteness (kriteria dan penjelasan), activation (mis. keberadaan pertanyaan pemandu), empathy (nada, penggunaan orang pertama), specificity, serta errors (kesalahan isi). Analisis statistik dilakukan untuk membandingkan LLM vs novice vs expert, dengan pengujian post hoc (Bonferroni) pada beberapa subkategori. Dengan pendekatan ini, dilakukan penilaian tidak hanya pada "siapa yang lebih baik", tetapi juga pada aspek apa LLM unggul/kurang, sehingga relevan bagi desain intervensi dan prompting ke depan.

Studi tersebut menunjukkan bahwa hanya prompt berkualitas tinggi yang secara konsisten memicu LLM menghasilkan umpan balik bermutu, sehingga desain prompt menjadi faktor kunci. Dibanding penilai pemula, LLM unggul pada hampir semua subkategori, sementara pada aspek afektif tertentu seperti valensi dan penggunaan orang pertama keunggulan itu tidak tampak. Dibanding pakar, LLM melampaui pada explanation, questions, dan specificity dengan selisih rata-rata (M_{diff}) masing-masing 0,46 (CI 95%:

0,17–0,74; p<0,001), 0,50 (CI 95%: 0,07–0,93; p<0,05), dan 0,96 (CI 95%: 0,52–1,41). Dari sisi efisiensi, ChatGPT-4 mampu menghasilkan sekitar 49 umpan balik dalam waktu yang sama ketika seorang pakar hanya menyusun satu, sehingga LLM berpotensi menjadi alternatif yang skalabel untuk umpan balik berkualitas asalkan perancangan prompt dilakukan dengan tepat.

Dilakukan penekanan bahwa kualitas prompt harus diajarkan sebagai keterampilan bagi pendidik; manual yang dikembangkan dapat dijadikan panduan praktis. Namun, terdapat keterbatasan yang perlu diperhatikan misalnya, halusinasi masih mungkin terjadi pada prompt tertentu, dan persepsi siswa terhadap umpan balik LLM serta transfer ke konteks/mata kuliah lain perlu diteliti lebih lanjut. Dengan kata lain, LLM berpotensi menjadi alternatif berkualitas dan efisien untuk sebagian peran feedback pakar, tetapi desain prompt serta pengawasan pedagogis tetap penting agar umpan balik selaras dengan kebutuhan dan konteks kursus.

2.1.3 *Reinforcement Learning from Human Feedback* (RLHF) untuk Peningkatan Kualitas Respons

Makalah berjudul "*Training language models to follow instructions with human feedback*" ini ditulis oleh Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, dkk. [18], sebagai proyek tim Alignment di OpenAI. Mereka memulai dari pengamatan bahwa memperbesar model bahasa tidak otomatis membuatnya lebih mampu mengikuti niat pengguna; model besar tetap bisa menghasilkan keluaran yang tidak benar, toksik, atau tidak membantu sehingga tidak selaras (misaligned) dengan kebutuhan pengguna. Paper ini menawarkan arah untuk menyelaraskan model dengan niat pengguna pada beragam tugas melalui penyelarasan berbasis umpan balik manusia (fine-tuning with human feedback).

Masalah utamanya adalah ketidakselarasan tujuan pelatihan model bahasa (memprediksi token berikutnya dari web) dengan tujuan penggunaan di dunia nyata (mengikuti instruksi secara membantu, jujur, dan aman). Konsekuensinya, meski bisa diprompt untuk banyak tugas, model sering "membuat fakta", bias/toksik, atau tidak mengikuti instruksi pengguna. Tim menekankan kebutuhan menyelaraskan model agar helpful-honest-harmless dalam penggunaan sehari-hari. Solusi mereka adalah pipeline RLHF (reinforcement learning from human feedback) dengan tiga tahap: (1) Supervised fine-tuning (SFT) pada demonstrasi yang ditulis pelabel tentang perilaku model yang diinginkan; (2) pelatihan reward model (RM) dari perbandingan/ranking beberapa keluaran model oleh pelabel; (3) optimasi kebijakan dengan PPO menggunakan RM sebagai sinyal ganjaran. Dataset dikumpulkan dari prompt yang ditulis pelabel dan prompt nyata yang disubmit melalui OpenAI API, sehingga mencerminkan distribusi penggunaan riil. Hasilnya adalah keluarga model InstructGPT.

Secara operasional, RM akhir diinisialisasi dari GPT-3 6B yang telah di-*fine-tune* pada beragam dataset NLP publik; pelatihan RM dilakukan satu *epoch* pada himpunan latih, dan mereka menemukan pelatihan sensitif terhadap jumlah *epoch* (multi-*epoch* cepat *overfit*), sementara *learning rate* tidak terlalu sensitif. Untuk model RLHF, kebijakan diinisialisasi dari model SFT dengan campuran 10% data pra-latih (membantu stabilitas PPO), kemudian dilatih sebanyak 256,000 episode yang mencakup sekitar 31,000 *prompt* unik (setelah penyaringan PII dan deduplikasi). *Batch* per iterasi sebanyak 512 (dipecah menjadi 8 *minibatch* berukuran 64) dan hanya satu *inner epoch* per iterasi; *warmup* pada 10 iterasi pertama, ditambah KL reward mengikuti Stiennon dkk. ($\beta = 0{,}02$). Rincian ini memperlihatkan setup yang hati-hati agar PPO stabil pada skala 1,3B/6B/175B.

Evaluasi utama menggunakan preferensi manusia di atas distribusi prompt API (mencerminkan use-case riil), serta metrik safety seperti truthfulness dan toxicity (dengan regresi kinerja minimal pada dataset NLP publik). Mereka juga memformalkan tujuan evaluasi selaras dengan kerangka *helpful, honest, harmless*. Visualisasi utama (Figur 1) menampilkan seberapa sering keluaran model lebih disukai dibanding *baseline* GPT-3 175B (SFT).

Secara konsisten, InstructGPT menunjukkan perbaikan yang berarti: model 1.3B mereka lebih sering dipilih manusia dibanding GPT-3 175B, meski 100 kali lebih kecil; ada peningkatan truthfulness dan penurunan toksisitas, sambil menjaga regresi kinerja pada dataset publik tetap minimal. Ini memperlihatkan bahwa penyelarasan dengan umpan balik manusia lebih menentukan daripada sekadar memperbesar parameter. Selain itu terdapat keterbatasan model yang masih bisa melakukan kesalahan sederhana. Namun, arah ini dinilai menjanjikan untuk menyelaraskan model dengan niat pengguna di banyak aplikasi.

2.1.4 Regulasi Diri dalam Pendidikan Pemrograman Berbasis LLM dengan *Learning Analytics*

Makalah *Design of AI-Powered Tool for Self-Regulation Support in Programming Education* karya Huiyong Li dan Boxuan Ma dari Kyushu University [19], disajikan pada CHI 2025 Workshop: Augmented Educators and AI, berangkat dari tantangan integrasi AI di pendidikan pemrograman. Banyak alat berbasis LLM yang digunakan mahasiswa berjalan di luar LMS institusional seperti Moodle. Keterpisahan ini memutus keterkaitan penting dengan konteks perkuliahan, misalnya materi kuliah, rincian soal, dan hasil eksekusi kode. Dampaknya, umpan balik AI sering kurang selaras dengan tujuan dan ruang lingkup mata kuliah, sementara dosen kesulitan menelusuri dampak jangka panjang interaksi mahasiswa dengan AI karena tidak ada jejak data proses yang terkumpul secara sistematis di lingkungan LMS. Pada saat yang sama, sebagian besar riset LLM di domain ini masih memfokuskan akuisisi pengetahuan (seperti generasi kode dan perbaikan bug)

sehingga perhatian pada keterampilan SRL seperti perencanaan, monitoring, dan refleksi—relatif kurang. Kekhawatiran lain adalah ketergantungan mahasiswa pada jawaban instan LLM, yang berisiko mengurangi upaya kognitif mendalam yang dibutuhkan pemula untuk membangun kemandirian belajar.

Untuk menjawab celah tersebut, dirancang solusi bernama CodeRunner Agent, sebuah alat LLM yang menempel langsung pada ekosistem Moodle. Ia memadukan *lecture viewer* untuk penyajian materi, plugin CodeRunner guna menjalankan dan menilai kode secara otomatis, serta *xAPI logging* ke Learning Record Store sehingga setiap interaksi penting—mulai dari akses materi, percobaan *test case*, hingga permintaan bantuan yang terekam sebagai data proses. Di sisi pedagogis, dukungan SRL dioperasionalkan lewat model PPESS (*Planning, Program creation, Error correction, Self-monitoring, Self-reflection*) yang menurunkan kerangka SRL Zimmerman khusus untuk pembelajaran pemrograman. Dengan pendekatan ini, mahasiswa bisa meminta bantuan sesuai fase SRL yang sedang dijalani, sehingga umpan balik AI tidak acak, melainkan terarah pada kebutuhan regulasi diri saat itu.

Arsitektur CodeRunner Agent menonjol karena mengandalkan dua mesin konteks. Pertama, LACE (*Learning Analytics Context Engine*) yang merangkum metrik keterlibatan dan performansi dari jejak xAPI, seperti waktu yang dihabiskan, frekuensi upaya, tingkat keberhasilan, dan pola kesalahan, lalu menyuntikkan ringkasan tersebut ke *prompt* agar respons LLM memanfaatkan sinyal perilaku SRL nyata. Kedua, KCE (*Knowledge Context Engine*) yang mengelola basis pengetahuan materi kuliah dan latihan, termasuk konsep kunci, prasyarat, tingkat kesulitan, solusi, dan *typical mistakes*, sehingga umpan balik AI selaras kurikulum dan menghindari kebocoran solusi langsung. Dosen dapat mengunggah materi, memutakhirkan basis pengetahuan, dan menyetel parameter konteks agar gaya dan standar umpan balik "menjadi milik mata kuliah" alih-alih generik.

Skenario penggunaan disiapkan untuk kelas pemrograman pemula. Mahasiswa menulis kode di CodeRunner, menjalankan *test cases*, lalu mengakses panel LLM-based Support dan memilih fase SRL yang relevan, misalnya *error correction* ketika menghadapi pesan kesalahan. Permintaan itu seperti identitas latihan, waktu, dan tipe bantuan tercatat sebagai xAPI. Dengan demikian terbentuk alur *end-to-end* dari log pembelajaran, rekayasa konteks, dan respons LLM yang dapat dianalisis untuk penyelarasan pedagogis dan perbaikan berkelanjutan. Pada tahap ini, makalah berposisi sebagai paper rancangan dengan merencanakan *pilot* jangka pendek dan eksperimen satu semester lintas beberapa kelas untuk memvalidasi rancangan pada situasi nyata.

Rencana evaluasi memanfaatkan log xAPI untuk melihat perubahan pola percobaan, akurasi, waktu penyelesaian, serta pergeseran strategi SRL ketika dukungan AI diaktifkan. Di luar metrik kuantitatif, dirancang juga studi kelas guna menilai kuali-

tas, relevansi, dan keamanan pedagogis dari umpan balik LLM, termasuk apakah agen berhasil menjaga SRL (memberi *scaffold* dan pertanyaan pemandu) alih-alih langsung membocorkan jawaban. Pendekatan evaluasi ganda ini menggabungkan analitik perilaku yang objektif dan penilaian pedagogis yang kualitatif ditujukan untuk memastikan alat tidak hanya berfungsi, tetapi bermanfaat bagi kemandirian belajar mahasiswa.

Kontribusi utama paper ini adalah rancangan terintegrasi yang menghubungkan data proses LMS dengan rekayasa konteks dan *scaffolding* SRL berbasis LLM di lingkungan yang benar-benar digunakan kampus. Pemetaan PPESS memberi kerangka operasional agar bantuan AI selalu terkait fase SRL yang tepat, sementara LACE dan KCE memastikan umpan balik kontekstual, selaras kurikulum, dan mendorong kemandirian. Hasil yang ditekankan bukan angka evaluasi, melainkan kelayakan implementasi dan janji untuk kelas besar yang ingin memanfaatkan analitik pembelajaran serta dukungan AI tanpa mengorbankan tujuan pengembangan SRL. Validasi empiris disiapkan melalui *pilot* dan studi semeseter penuh pada tahap berikutnya.

2.1.5 Dashboard *Real-time* untuk Orkestrasi Kelas Berbasis Kanban

Makalah ini ditulis oleh Sven Strickroth, Melanie Kreidenweis, dan Andreas Götzfried. Mereka menyoroti tantangan nyata pada pembelajaran berbasis tugas di kelas yang bersifat kolaboratif. Ketika beberapa kelompok bergerak pada tugas berbeda dalam waktu yang sama, guru sulit memiliki pandangan menyeluruh tentang status pekerjaan setiap kelompok. Versi awal alat mereka, AgileBoard4Teaching, memang memungkinkan perencanaan dan pelaksanaan dengan papan Kanban per kelompok, tetapi berjalan sepenuhnya *offline*. Guru perlu menyalurkan berkas ke semua siswa, berkeliling memeriksa satu per satu, serta tidak memiliki ringkasan progres kelas secara langsung. Di sisi lain, solusi profesional seperti Trello tidak ideal untuk sekolah karena fitur dan kebijakan privasi. Kebutuhan akan *dashboard real-time* menjadi titik mula rancangan mereka.

Solusi yang diusulkan adalah memperluas AgileBoard4Teaching menjadi sistem berjaringan dengan arsitektur *client-server*. Di sisi guru tersedia *authoring mode* untuk menyusun tugas dan putaran kerja lalu menggandakan papan untuk banyak kelompok. Di sisi siswa tersedia antarmuka Kanban tiga kolom *ToDo*, *In Progress*, *Done* dengan unggah hasil dan permintaan *review*. Inti pembeda adalah *dashboard* guru yang menampilkan kotak ringkas untuk tiap kelompok. Guru dapat melihat jumlah tugas per kolom, daftar pekerjaan yang menunggu persetujuan, serta sinyal seperti *raise hand*. Tersedia pula pengiriman pesan massal, penetapan *timer*, dan jeda tampilan siswa. Secara teknis, sistem menggunakan Java Servlets dan MariaDB di *backend*, penyimpanan papan sebagai JSON, komunikasi WebSocket untuk sinkronisasi perubahan secara langsung, dan opsi *deployment* via Docker. Mekanisme bergabung untuk siswa menggunakan kode grup agar tetap ramah privasi tanpa akun individual.

Dilakukan dua studi evaluasi. Pertama, studi lapangan di kelas TIK tingkat 8 pada topik *CAD* selama tiga minggu dengan dua sesi per minggu. Desainnya *within-subjects*. Satu minggu pertama menggunakan versi *offline*, lalu beralih ke versi jaringan untuk membandingkan pengalaman secara adil. Total 18 siswa dalam 9 kelompok berpartisipasi. Pengumpulan data meliputi observasi langsung, kuesioner fitur, *System Usability Scale* (SUS), wawancara dengan guru, serta analisis *log* interaksi server. Kedua, studi simulasi berupa lokakarya dua jam yang diikuti 18 guru dan *trainer* berpengalaman. Pada lokakarya ini, prototipe dipakai untuk mengorkestrasi kegiatan sehingga peserta mengalami peran "siswa" sekaligus menilai kegunaan alat.

Temuan dari studi lapangan menunjukkan pergeseran preferensi yang kuat ke versi jaringan. Siswa menyatakan alur kerja lebih lancar karena tidak perlu impor dan simpan manual. Fitur seperti *raise hand*, pesan, *timer*, dan *pause* dinilai membantu koordinasi. SUS siswa mencapai 84 yang tergolong sangat baik dan SUS guru 92 yang termasuk kategori pengalaman terbaik. Analisis *log* memperlihatkan intensitas penggunaan tinggi di kelas serta pemanfaatan di luar kelas untuk persiapan dan *review*. Guru menyampaikan bahwa *dashboard* paling berguna untuk meninjau progres banyak kelompok sekaligus dan menyiapkan pertemuan berikutnya. Selama sesi tatap muka, guru tidak selalu menatap *dashboard* karena fokus pada pendampingan langsung. Beberapa saran muncul seperti tampilan yang lebih ringkas untuk kelas besar, pembuatan putaran baru secara cepat, serta statistik waktu dan progres.

Pada studi simulasi, para guru mengapresiasi sinkronisasi waktu nyata dan kemudahan kloning papan untuk banyak kelompok. SUS rata-rata 69 yang tetap masuk kategori baik. Skor ini dipengaruhi kejadian *bug* yang membuat sebagian peserta terkendala masuk ke putaran kedua. Meski demikian, mayoritas peserta menyatakan ingin mencoba alat ini di kelas mereka. Masukan tambahan menyangkut kebutuhan versi terkelola yang stabil dan peningkatan visibilitas opsi komunikasi.

Secara keseluruhan, kontribusi utama makalah adalah desain dan prototipe *dashboard real-time* yang mendukung orkestrasi kelas *agile* berbasis Kanban, lengkap dengan bukti kegunaan dari dua konteks evaluasi. Solusi ini menjembatani celah antara alat perencanaan yang kuat dan kebutuhan pemantauan langsung di kelas. Hasilnya menyiratkan kesiapan untuk dilanjutkan ke tahap produksi dan pengayaan fitur *learning analytics* seperti waktu penyelesaian dan deviasi dari target. Ditekankan juga pentingnya menjaga keseimbangan antara informasi yang cukup dan beban kognitif guru agar *dashboard* tetap dapat dibaca sekejap sekaligus memberi sinyal tindakan yang jelas.

2.1.6 Perbandingan Metode Peningkatan Kualitas Respons LLM

Dapat dilihat, Tabel 2.1 merangkum karya kunci yang mewakili dua arus utama yakni penyelarasan *model* melalui *fine-tuning* misalnya *DPO* dan *RLHF* serta penye-

larasan *konteks* melalui *prompt* dan *context engineering* serta orkestrasi kelas berbasis Kanban.

Tabel 2.1. Komparasi penelitian peningkatan kualitas umpan balik LLM di pendidikan

Studi	**Tujuan**	**Pendekatan Inti**	**Pelaksanaan Penelitian**	**Temuan Utama**
Woodrow et al. (EDM 2025) [1]	Menyelaraskan gaya dan isi feedback AI dengan preferensi pengajar	*Direct Preference Optimization* dengan guru/TA di dalam loop. Preferensi menang-kalah dari dua kandidat umpan balik dipakai untuk fine-tuning	Mata kuliah besar. Siklus antartugas. Observasi pekerjaan mahasiswa diringkas ke prompt terstruktur	Studi terkontrol menunjukkan DPO lebih disukai daripada GPT-4o. Skor insight lebih tinggi. *Naturalness* setara. Evaluasi otomatis juga lebih baik pada empat dari lima dimensi. DPO kadang lebih panjang
Jacobsen & Weber (2025) [17]	Menentukan desain prompt yang memicu feedback LLM berkualitas serta membandingkan LLM dengan manusia	Manual prompt berbasis teori. Tiga kualitas prompt rendah menengah tinggi. LLM dibanding novice dan expert pada tugas geometri	Feedback pada learning goal dan miskonsepsi umum. Semua pihak memakai prompt berkualitas tinggi untuk pembandingan adil	Prompt berkualitas tinggi menghasilkan feedback terbaik. LLM mengungguli novice di hampir semua subkategori. LLM melampaui expert pada explanation questions specificity. Efisiensi LLM jauh lebih tinggi
Ouyang et al. (InstructGPT, 2022) [18]	Menyelaraskan model agar mengikuti instruksi yang membantu jujur dan aman	*RLHF* tiga tahap. SFT pada demonstrasi manusia. Reward model dari perbandingan keluaran. PPO untuk mengoptimalkan kebijakan	Berbagai tugas umum. Prompt dari pelabel dan pengguna API. Bukan konteks kursus spesifik	Model kecil yang telah selaras lebih sering dipilih manusia dibanding GPT-3 yang jauh lebih besar. Truthfulness naik dan toksisitas turun. Kerangka selaras tugas umum bukan feedback proses belajar
Li & Ma (CHI 2025 WS) [19]	Merancang alat LLM yang mendukung SRL di pemrograman serta tetap terhubung ke LMS	Rekayasa konteks ganda. LACE untuk ringkasan learning analytics berbasis xAPI. KCE untuk pengetahuan kurikulum. Pemetaan SRL ke PPESS	Moodle dengan CodeRunner. Log xAPI merekam akses materi upaya tes dan bantuan. Fase SRL dipilih pengguna untuk memicu bantuan yang tepat	Paper rancangan. Rencana uji lapangan dan studi satu semester. Menekankan kelayakan integrasi. Fokus pada dukungan SRL dan kontrol kebocoran jawaban. Belum ada angka hasil akhir
Strickroth et al. (2025) [20]	*Dashboard real-time* untuk orkestrasi kelas Kanban agar guru memantau progres banyak kelompok	Ekstensi AgileBoard4Teaching ke sistem *client–server*. *Dashboard* guru. Sinkronisasi *WebSocket*. Penyimpanan papan JSON. Opsi *deployment* Docker. Mekanisme kode grup ramah privasi	Kelas TIK tingkat 8 topik *CAD*. 18 siswa 9 kelompok. Desain *within-subjects offline* ke jaringan. Lokakarya dua jam dengan 18 guru/*trainer*. Data observasi kuesioner SUS wawancara dan *log* server	Preferensi kuat ke versi jaringan. Alur kerja lebih lancar. SUS siswa 84 dan SUS guru 92. Lokakarya SUS 69. Masukan lanjutan meliputi tampilan ringkas kelas besar pembuatan putaran cepat serta statistik waktu dan progres

Tabel 2.1 secara garis besar menjelaskan karya terdahulu yang berpencar ke dua arus utama. Arus pertama berfokus pada penyelarasan *model* melalui *fine-tuning*, misalnya *Direct Preference Optimization* (DPO) [1] dan *Reinforcement Learning from Human Feedback* (RLHF) [18]. Arus kedua menekankan penyelarasan *konteks* melalui perancangan *prompt* dan lingkungan interaksi beserta pemanfaatan data proses pembelajaran, sebagaimana tampak pada [17] dan [19]. Di luar itu, [20] menawarkan orkestrasi kelas berbasis papan Kanban dengan *dashboard real-time* yang kaya sinyal proses sehingga menyediakan landasan operasional untuk memantau dan menindaklanjuti progres kelompok secara langsung.

Dari sisi tujuan intervensi, DPO dan RLHF mengubah parameter model agar keluaran lebih selaras dengan preferensi manusia [1, 18]. Sebaliknya, [17] memperlihatkan bahwa *prompt* berkualitas tinggi dapat mendorong kualitas umpan balik tanpa perlu pelatihan ulang, sementara [19] menekankan rekayasa konteks yang ditopang oleh *learning analytics*. Penelitian Woodrow juga tidak berfokus pada konteks SRL, melainkan pada penyelarasan gaya dan isi umpan balik sesuai preferensi pengajar. Walaupun metrik evaluasi otomatis yang digunakan Woodrow dkk. (kerangka Scarlatos [16]) mendukung adanya fungsi pedagogis yang mendukung SRL, hal tersebut bukan fokus utama penelitian mereka.

Menariknya, [20] secara tidak langsung menyediakan fitur yang mendukung guru untuk melakukan tugas *learning analytics*. Hal ini dapat dilihat dari kemampuan *dashboard* dalam menampilkan seperti jumlah tugas per kolom, daftar pekerjaan, dan waktu yang dihabiskan oleh siswa pada setiap tugas. Dengan demikian, guru dapat memantau aktivitas pembelajaran secara real-time dan mengambil tindakan yang diperlukan. Penelitian ini memosisikan diri pada jalur kedua, yaitu optimalisasi konteks, sejalan dengan batasan metodologis bahwa tidak dilakukan *fine-tuning*.

Penelitian [17] memperlihatkan bagaimana rekayasa *prompt* (*prompt engineering*) memengaruhi kualitas umpan balik dari *Large Language Models* (LLM). Metode yang digunakan adalah mengembangkan sebuah panduan perancangan *prompt* berbasis (*theory-driven prompt manual*) untuk menciptakan tiga jenis *prompt* dengan kualitas berbeda yaitu rendah, sedang, dan tinggi. Berikut adalah hasil dari yang dikerjakan oleh Jacobsen.

Tabel 2.2. Prompt manual to ensure the development of high-quality prompts.

Kategori	**Subkategori**	**Tinggi (2)**	**Sedang (1)**	**Rendah (0)**
Context	Role	Peran LLM dan penanya dijelaskan	Hanya satu peran yang dijelaskan	Tidak ada peran LLM maupun penanya yang dijelaskan
	Target audience	Ada target audiens yang jelas didefinisikan dan dijelaskan	Target audiens dijelaskan secara kasar	Target audiens tidak ditentukan
	Medium/channel	Media atau saluran tempat informasi disajikan dijelaskan dengan jelas	Media atau saluran dijelaskan secara kasar	Media atau saluran tidak disebutkan
Mission	Mission/question	Misi LLM dijelaskan dengan jelas	Misi LLM dijelaskan secara kasar	Misi LLM tidak jelas
Clarity dan specificity	Format and constraints	Properti stilistika serta spesifikasi panjang yang dideskripsikan	Hanya properti stilistika ATAU spesifikasi panjang yang diberikan	Tidak ada properti stilistika maupun spesifikasi panjang yang diberikan
	Conciseness	*Prompt* hanya berisi informasi yang langsung terkait dan relevan dengan output, serta jelas dan ringkas	*Prompt* ringkas dengan sedikit informasi berlebih	*Prompt* mengandung banyak informasi yang tidak relevan dengan misi/pertanyaan
	Domain specificity	Istilah teknis digunakan dengan benar dan memungkinkan LLM merujuknya di dalam jawaban	Istilah teknis digunakan sesekali atau tanpa penjelasan	Tidak menggunakan kosakata spesifik yang relevan dengan bidang subjek
	Logic	*Prompt* memiliki alur baca yang sangat baik, koherensi logis internal, urutan informasi yang koheren, dan hubungan yang jelas antara konten dan misi	*Prompt* hanya memenuhi beberapa kondisi ini	*Prompt* disusun secara tidak logis

Studi ini menemukan bahwa hanya *prompt* dengan kualitas terbaik yang secara konsisten menghasilkan umpan balik berkualitas tinggi, yang kualitasnya bahkan melampaui umpan balik dari para ahli (*experts*) dalam beberapa kategori seperti penjelasan, pertanyaan, dan spesifisitas. Pendekatan ini sangat berpotensi untuk diterapkan karena menawarkan efisiensi tinggi dengan menghasilkan umpan balik berkualitas yang jauh lebih cepat daripada ahli. Hal ini juga tidak memerlukan *fine-tuning* model, sehingga relatif murah dari segi sumber daya komputasi dan waktu.

Untuk perspektif sumber daya dan kelayakan, pendekatan DPO/RLHF menuntut data preferensi atau label serta infrastruktur pelatihan. Hal tersebut tidak selaras dengan rancangan studi ini yang memanfaatkan model *open source* melalui Groq API tanpa pe-

latihan ulang. Pendekatan *prompt/context engineering* lebih ringan dijalankan, mudah direplikasi, dan sesuai dengan keterbatasan komputasi maupun waktu. Penelitian [20] memanfaatkan *dashboard real-time* dan jejak proses Kanban yang memungkinkan pemantauan banyak kelompok sekaligus, pengumpulan sinyal proses yang autentik, dan pengambilan tindakan cepat oleh pengajar. Tidak seperti penelitian Li dan Ma [19] yang menggunakan Moodle sebagai *backbone* platform, platform *dashboard real-time* milik Strickroth dkk. dapat memberikan data yang lebih luas dan fleksibel untuk memantau keseluruhan proses pembelajaran kolaboratif berbasis Kanban. Penggunaan papan Kanban oleh pada penelitian [20] juga sesuai dengan keselarasan pada pembahasan latar belakang penelitian ini yaitu pemanfaatan jejak proses belajar mahasiswa yang direpresentasikan dalam bentuk papan kanban. Komponen-komponen ini belum tercakup pada studi yang berfokus pada *prompt* semata oleh Ouyang [18]. Hal ini membuat beberapa komponen yang ada di penelitian Strickroth dkk. dapat dimanfaatkan untuk memperkaya konteks umpan balik LLM.

Dari sisi kontrol pedagogis, tetap dipertahankan melalui penilaian kualitatif yang akan ditinjau langsung oleh pakar pendidikan. Keluaran LLM yang berupa *feedback* ditinjau dan dinilai melalui rubrik penilaian yang telah dipersiapkan. Dengan cara ini, kualitas pedagogis tetap terjaga dan sejalan dengan kebutuhan untuk dilakukan penilaian oleh pakar manusia.

Dengan demikian, penelitian ini mendapatkan celah perbaruan di mana pendekatan *prompting* dan *context engineering* yang murah sumber daya dan biaya digabungkan dengan *learning analytics* yang berasal dari papan Kanban untuk memperkaya konteks umpan balik LLM. Kombinasi tersebut belum pernah dibahas sebelumnya oleh literatur yang ada. Celah penelitian ini akan dievaluasi secara kuantitatif untuk mengukur kualitas keluaran LLM terhadap acuan pakar serta secara kualitatif melalui penilaian pakar pendidikan untuk menjaga kualitas pedagogis. Melalui batasan-batasan yang telah ditetapkan, rancangan ini menawarkan solusi praktis dan efektif. Pilihan metodologis ini konsisten dengan batasan penelitian, memanfaatkan data proses untuk memperkuat konteks *prompt*, menjaga peran guru sebagai pengarah kualitas, dan menghadirkan kesiapan implementasi di lapangan.

2.1.7 Metode Evaluasi Umpan Balik LLM dalam Pendidikan

Evaluasi kualitas teks yang dihasilkan oleh *Large Language Models* (LLM) merupakan tantangan kompleks yang memerlukan pendekatan multi-aspek. Secara historis, metrik evaluasi didominasi oleh metode berbasis tumpang tindih leksikal seperti BLEU dan ROUGE. Metrik ini bekerja dengan menghitung kesamaan *n-gram* (urutan kata) antara teks yang dihasilkan dan teks referensi (*ground truth*) [21]. Namun, ketergantungan pada pencocokan permukaan ini memiliki kelemahan signifikan; metrik ini gagal mema-

hami kesetaraan makna (semantik) dan sering kali memberikan skor rendah pada parafrasa yang valid, ketika dua kalimat memiliki arti yang sama tetapi menggunakan kosakata yang berbeda [22]. Berbagai studi telah menunjukkan bahwa korelasi metrik ini dengan penilaian manusia sering kali rendah, terutama dalam tugas-tugas yang menuntut keragaman ekspresi seperti simplifikasi teks atau dialog [23].

Keterbatasan ini mendorong pergeseran paradigma menuju metrik evaluasi yang mampu menangkap kesamaan semantik. Pada penelitian ini, evaluasi yang akan dilakukan dibatasi pada dua faktor yaitu kesamaan semantik dan kualitas relevansinya terhadap konteks dan kriteria tertentu. Untuk pendekatan kesamaan semantik, metrik berbasis *embedding* kontekstual seperti BERTScore dan BARTScore menjadi pilihan utama [24, 25]. Lalu, penilai (dikenal sebagai *LLM-as-a-Judge*) dapat memberikan penilaian relevansinya terhadap kriteria tertentu untuk teks yang dihasilkan. Metrik berbasis *embedding* seperti BERTScore dan BARTScore tidak lagi membandingkan kata secara literal, melainkan representasi vektornya yang kaya akan konteks, sehingga lebih selaras dengan penilaian kualitas oleh manusia [24]. Di sisi lain, pendekatan *LLM-as-a-Judge* memanfaatkan kemampuan pemahaman bahasa dari model canggih seperti GPT-4 untuk memberikan penilaian kualitatif yang bernuansa, meniru proses evaluasi oleh manusia pada skala yang lebih besar [26].

Dalam paper ini [24] Metrik ini bekerja dengan mengubah setiap token dalam kalimat kandidat dan referensi menjadi *embedding* kontekstual menggunakan model BERT. Selanjutnya, kesamaan makna diukur dengan menghitung kemiripan kosinus (*cosine similarity*) antara setiap pasang token, lalu secara serakah (*greedy matching*) mencocokkan token dari satu kalimat ke token yang paling mirip di kalimat lainnya untuk menghasilkan skor presisi, *recall*, dan F1. Keunggulan utamanya adalah kemampuannya untuk mengenali sinonim dan parafrasa, sehingga memberikan evaluasi yang lebih kuat terhadap kesetaraan makna dibandingkan metrik *n-gram*.

Sementara itu, BARTScore [25] menawarkan filosofi yang berbeda secara fundamental dengan membingkai evaluasi sebagai "tugas generasi teks". Alih-alih hanya membandingkan *embedding*, BARTScore menggunakan model seq2seq (sequence-to-sequence) untuk menghitung probabilitas sebuah teks dihasilkan dari teks lainnya. Misalnya, untuk mengukur presisi, ia akan menghitung seberapa mungkin model BART akan menghasilkan teks referensi jika diberi teks kandidat sebagai sumber. Skor yang lebih tinggi (*log-probability* yang lebih mendekati nol) menandakan kualitas yang lebih baik. Fleksibilitas ini memungkinkan BARTScore untuk menilai berbagai aspek seperti kefasihan, koherensi, dan bahkan faktualitas hanya dengan mengubah input sumber dan targetnya.

Untuk mengatasi subjektivitas dan bias dalam pendekatan *LLM-as-a-Judge*, kerangka kerja terstruktur seperti G-Eval telah dikembangkan [26]. G-Eval menggunakan

LLM penilai (misalnya, GPT-4) tetapi memandunya dengan penalaran *Chain-of-Thought* (CoT). Alih-alih hanya meminta skor akhir, G-Eval terlebih dahulu menginstruksikan LLM untuk menghasilkan langkah-langkah evaluasi berdasarkan kriteria yang diberikan pengguna, lalu mengikuti langkah-langkah tersebut untuk sampai pada sebuah skor. Pendekatan ini membuat proses penilaian lebih transparan, dapat dikontrol, dan terbukti mencapai korelasi yang lebih tinggi dengan penilaian manusia dibandingkan metode LLM-penilai yang tidak terstruktur. Pada penelitian [27], dilakukan juga penilaian umpan balik kode pemrograman yang dihasilkan oleh model open-source. Seperti G-Eval, penilai LLM diminta untuk menilai berdasarkan kriteria yang telah ditentukan. Sehingga dapat disimpulkan bahwa pendekatan *LLM-as-a-Judge* yang terstruktur dengan kriteria jelas meningkatkan keandalan penilaian.

2.1.8 Evaluasi Ahli Manusia

Untuk menilai kualitas umpan balik yang dihasilkan oleh LLM secara komprehensif, evaluasi manusia yang terstruktur menjadi standar emas, karena mampu menangkap nuansa pedagogis dan psikologis yang terlewat oleh metrik otomatis. Penilaian ini dapat dibingkai menggunakan rubrik yang didasarkan pada prinsip-prinsip teoretis yang telah mapan. Mengacu pada Shute (2008) [28], kriteria fundamental yang harus dinilai adalah apakah umpan balik LLM bersifat jelas, akurat, dan relevan dengan konteks tugas yang diberikan. Penilai manusia dapat secara langsung memverifikasi apakah instruksi yang diberikan spesifik dan apakah informasi faktualnya benar.

Selanjutnya, dengan mengadopsi model dari Hattie dan Timperley (2007) [29], penilai dapat mengevaluasi kelengkapan struktural umpan balik tersebut. Rubrik penilaian akan memeriksa apakah LLM mampu menjawab tiga pertanyaan kunci: (1) Feed Up: Apakah umpan balik mengklarifikasi tujuan akhir (Where am I going?); (2) Feedback: Apakah ia memberikan analisis tentang kinerja saat ini terhadap tujuan tersebut (How am I going?); dan (3) Feed Forward: Apakah ia menawarkan saran yang konkret dan dapat ditindaklanjuti untuk perbaikan (Where to next?).

Terakhir, dampak motivasional dari umpan balik, sebuah aspek yang sulit diukur secara otomatis, dapat dinilai melalui lensa Self-Determination Theory (Deci & Ryan, 1985) [30] dan prinsip penguatan positif (Fredrickson, 2001) [31]. Penilai manusia akan mengevaluasi apakah bahasa yang digunakan LLM mendukung otonomi (memberikan pilihan, bukan mendikte), menumbuhkan rasa kompetensi (mengakui usaha dan kemajuan), dan menggunakan nada yang apresiatif untuk meningkatkan self-efficacy. Dengan demikian, evaluasi manusia yang berlandaskan teori ini tidak hanya mengukur kebenaran teknis dari umpan balik, tetapi juga efektivitasnya dalam memotivasi dan memberdayakan pengguna secara konstruktif.

Secara keseluruhan, penilaian kualitas umpan balik LLM oleh ahli manusia meng-

gabungkan evaluasi objektif berdasarkan rubrik yang dikategorikan menjadi tiga jenis yaitu penilaian secara *feedback*, *motivation support*, dan *appreciation support*. Setiap kategori dievaluasi menggunakan skala Likert (1-5) untuk mengukur berbagai dimensi kualitas umpan balik dari perspektif pedagogis dan psikologis. Dengan cara ini, evaluasi manusia menyediakan gambaran yang kaya dan bernuansa tentang efektivitas umpan balik LLM dalam konteks pendidikan.

2.2 Pertanyaan Penelitian

Berdasarkan tujuan penelitian yang telah diuraikan dan metode evaluasi yang dipilih, serta mempertimbangkan sifat penelitian yang cenderung **eksploratif**, maka fokus utama penelitian ini dirumuskan ke dalam beberapa **pertanyaan penelitian** (*Research Questions*/RQ) sebagai berikut:

1. **(RQ 1)** Sejauh mana metode *context engineering* yang digabungkan dengan *learning analytics* (LA) dapat meningkatkan kualitas umpan balik LLM, jika dibandingkan dengan metode *Baseline*?
2. **(RQ 2)** Bagaimana persepsi dan penilaian ahli manusia terhadap kualitas, relevansi, dan kelayakan umpan balik yang dihasilkan oleh metode *context engineering* dengan LA?
3. **(RQ 3)** Sejauh mana ukuran model bahasa (misalnya, model kecil vs. model besar) memengaruhi kualitas umpan balik yang dihasilkan dalam konteks penelitian ini?

2.3 Dasar Teori

2.4 Dasar Teori

Bab ini menyajikan landasan konseptual dan kerangka teoretis yang menopang rancangan sistem, metodologi eksperimen, serta evaluasi penelitian. Pembahasan disusun secara sistematis mulai dari fondasi pedagogis hingga implementasi teknis dan metode pengujian.

Pertama, diuraikan **Tinjauan Teoretis Psikologi Pendidikan** yang menjadi standar acuan kualitas sistem, mencakup prinsip *Self-Regulated Learning* (SRL) dalam lingkungan berbasis Kanban , teori umpan balik formatif, serta kerangka motivasi *Self-Determination Theory* (SDT) dan psikologi positif. Kedua, dibahas karakteristik **Large Language Models (LLM)** sebagai mesin utama agen cerdas, dengan fokus pada arsitektur model Llama dan implikasi teoritis perbedaan ukuran parameter (8B vs 70B) terhadap kapasitas penalaran model. Ketiga, dijelaskan strategi **Rekayasa Konteks (*Context Engineering*)** dan *Prompt Engineering* sebagai jembatan antara pedagogi dan teknologi, termasuk penerapan mekanisme *ReAct (Reasoning + Acting)* untuk meningkatkan logika internal model. Keempat, dipaparkan pendekatan **Learning Analytics** (LA) tipe deskriptif yang digunak-

an untuk mentransformasi data mentah aktivitas Kanban menjadi wawasan kontekstual bagi LLM. Terakhir, bab ini menjelaskan metodologi **Evaluasi** yang menggabungkan metrik kesamaan semantik (*BERTScore*, *BARTScore*) dan pendekatan *LLM-as-a-Judge* untuk mengukur kualitas respons secara objektif.

2.4.1 Tinjauan Teoritis Psikologi Pendidikan

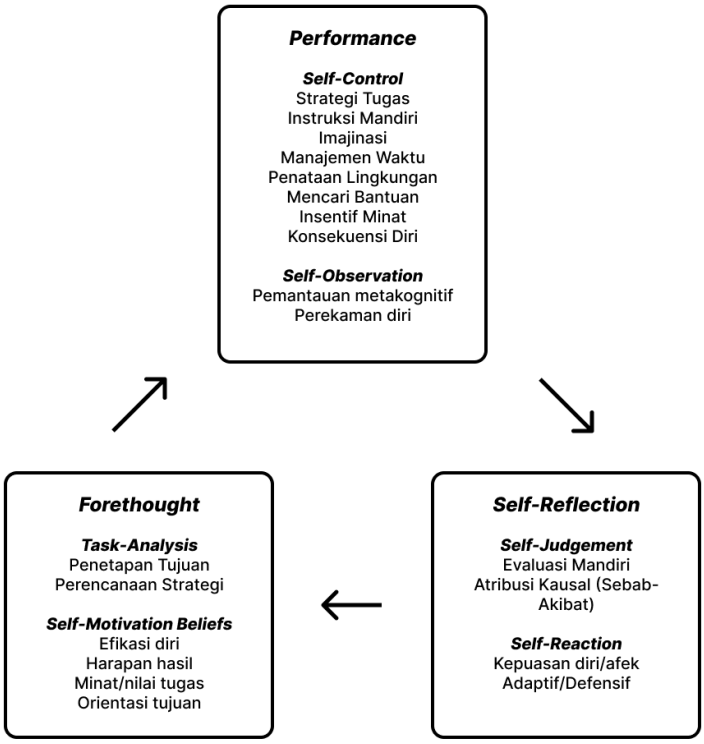
Pada bagian ini dibahas teori-teori psikologi pendidikan yang menjadi landasan konseptual bagi penelitian ini, meliputi *Self-Regulated Learning* (SRL), teori umpan balik formatif, *Self-Determination Theory* (SDT), dan *Positive Psychology*. Hal ini penting untuk memahami bagaimana umpan balik yang dihasilkan oleh LLM dapat dirancang agar efektif mendukung proses belajar mandiri mahasiswa. Selain itu, teori-teori ini memberikan kerangka untuk mengevaluasi kualitas umpan balik dari perspektif pedagogis yang akan dilakukan oleh ahli manusia pada bab evaluasi.

2.4.1.1 *Self-Regulated Learning* (SRL)

Self-Regulated Learning (SRL), atau Pembelajaran yang Diregulasi Diri, merujuk pada otonomi dan kontrol yang dilakukan oleh individu, yang secara aktif memantau, mengarahkan, dan meregulasi tindakan untuk mencapai tujuan seperti perolehan informasi, peningkatan keahlian, dan pengembangan diri [32]. Dijelaskan juga oleh Zimmerman [33] definisi dari SRL bahwa kondisi pelajar yang aktif baik secara metakognitif, motivasional, maupun perilaku dalam pembelajaran mereka. Secara metakognitif, pelajar terampil dalam merencanakan, mengorganisasi, menginstruksi diri, memantau diri, dan mengevaluasi diri pada berbagai tahap proses belajar. Secara motivasional, pelajar memiliki pandangan diri sebagai individu yang otonom, kompeten, dan yakin terhadap kemampuan mereka untuk berprestasi. Sementara itu, dari sisi perilaku, pelajar mampu memilih, menyusun, dan menciptakan lingkungan yang dapat mengoptimalkan proses belajar [33].

Untuk mencapai tujuan pembelajaran, SRL memiliki banyak model teoretis. Salah satu model yang paling banyak digunakan adalah model siklus tiga fase dari Zimmerman dan Moylan [2], melalui siklus perencanaan (*forethought*), pelaksanaan (*performance*), dan refleksi (*self-reflection*) yang saling terkait. Dalam fase *forethought*, pelajar menganalisis tugas, menetapkan tujuan, dan merencanakan strategi untuk mencapai tujuan tersebut, yang didorong oleh keyakinan motivasional. Fase *performance* melibatkan eksekusi tugas yang sebenarnya, di mana pelajar memantau kemajuan mereka dan menggunakan strategi kontrol diri untuk mempertahankan keterlibatan kognitif dan motivasi hingga tugas selesai. Terakhir, fase *self-reflection* adalah saat pelajar menilai kinerja mereka dan membuat atribusi terhadap keberhasilan atau kegagalan (misalnya melalui penilaian diri formatif), yang kemudian menghasilkan reaksi diri yang akan memengaruhi

bagaimana pelajar mendekati tugas serupa di masa mendatang. Gambar 2.2 mengilustrasikan model siklus tiga fase SRL ini.



Gambar 2.2. Model Siklus Tiga Fase *Self-Regulated Learning* (SRL) menurut Zimmerman dan Moylan [2].

Di konteks pendidikan tinggi, SRL berperan krusial karena lingkungan belajar menuntut kemandirian, terutama pada pembelajaran digital (daring atau campuran) yang memerlukan disiplin diri lebih tinggi serta kemampuan mengelola distraksi. Kajian terkini menunjukkan bahwa SRL berkorelasi dengan kinerja akademik, keterlibatan, dan ketahanan studi. Sebaliknya, kekurangan SRL di pembelajaran daring atau campuran dapat memperlebar kesenjangan capaian yang diperoleh mahasiswa [34]. Namun, tanpa bimbingan yang baik, mahasiswa sering kesulitan menafsirkan pemahaman mereka dalam melakukan pemikiran reflektif tersebut [34]. Dengan demikian, pengembangan kompetensi SRL termasuk penetapan tujuan, strategi, monitoring, dan refleksi—menjadi prasyarat penting untuk keberhasilan mahasiswa di ekosistem belajar modern [34].

2.4.1.2 Teori Umpan Balik

Dalam konteks pendidikan, umpan balik didefinisikan sebagai informasi yang dikomunikasikan kepada pelajar dengan niat untuk memodifikasi pemikiran atau perilaku guna meningkatkan pembelajaran. Hattie dan Timperley memandang umpan balik sebagai konsekuensi dari kinerja yang disediakan oleh agen seperti guru atau teman sebaya

yang bertujuan untuk mengurangi kesenjangan antara pemahaman saat ini dan tujuan yang diinginkan [29]. Shute [28] memperluas perspektif ini dengan menekankan bahwa umpan balik formatif harus bersifat non evaluatif suportif tepat waktu dan spesifik agar efektif. Secara kognitif umpan balik berfungsi untuk mengurangi ketidakpastian mengenai seberapa baik kinerja mahasiswa dan mengurangi beban kognitif atau *cognitive load* terutama bagi mahasiswa baru atau yang mengalami kesulitan belajar.

Efektivitas umpan balik dalam menutup kesenjangan pembelajaran bergantung pada kemampuannya menjawab tiga pertanyaan fundamental. Pertanyaan pertama adalah "**Ke mana saya akan pergi**" yang berkaitan dengan penetapan tujuan yang jelas dan menantang. Pertanyaan kedua adalah "**Bagaimana saya menjalaninya**" yang memberikan informasi mengenai kemajuan siswa relatif terhadap tujuan tersebut. Pertanyaan ketiga adalah "**Langkah selanjutnya**" yang mengarahkan siswa pada aktivitas lanjutan untuk memperdalam pemahaman [29]. Shute menambahkan bahwa umpan balik yang efektif dapat memfasilitasi pergeseran orientasi tujuan siswa dari orientasi kinerja menjadi orientasi *learning orientation* yang mendorong kegigihan dan penguasaan keterampilan [28].

Hattie dan Timperley mengategorikan fokus umpan balik ke dalam empat tingkatan utama. Tingkat tugas atau *Task Level* berfokus pada kebenaran faktual dari kinerja dan sering disebut sebagai umpan balik korektif. Tingkat proses atau *Process Level* menyasar strategi kognitif yang mendasari penyelesaian tugas. Tingkat regulasi diri atau *Self Regulation Level* mendukung otonomi dan pemantauan diri siswa. Tingkat diri atau *Self Level* yang berupa pujian pribadi dinilai paling tidak efektif karena mengalihkan perhatian dari tugas [29]. Shute [28] mendukung pandangan ini dengan temuan bahwa umpan balik yang bersifat spesifik atau terelaborasi (*elaborated feedback*) umumnya lebih efektif daripada umpan balik yang hanya berupa verifikasi benar atau salah terutama untuk pembelajaran mendalam. Namun kompleksitas umpan balik harus dikelola agar tidak membebani siswa dengan informasi yang berlebihan.

2.4.1.3 *Self-Determination Theory* (SDT)

Teori Determinasi Diri atau *Self Determination Theory* dijelaskan sebagai pendekatan organismik terhadap motivasi yang **memandang manusia sebagai makhluk aktif dengan kecenderungan alamiah menuju pertumbuhan psikologis serta integrasi diri** [30]. Perspektif ini menekankan pentingnya pemenuhan **tiga kebutuhan psikologis dasar yang bersifat bawaan: kebutuhan akan kompetensi, kebutuhan akan otonomi (determinasi diri), dan kebutuhan akan keterhubungan interpersonal**. Kebutuhan akan kompetensi berkaitan dengan keinginan individu untuk merasa efektif dalam berinteraksi dengan lingkungan sedangkan kebutuhan akan determinasi diri mengacu pada keinginan individu untuk menjadi sumber kausalitas dari perilakunya sendiri tanpa pak-

saan dari kekuatan luar [30].

Dalam kerangka teori ini, motivasi intrinsik didefinisikan sebagai energi yang didasarkan pada kebutuhan bawaan untuk merasa kompeten dan menentukan nasib sendiri yang mendorong individu mencari tantangan optimal demi kepuasan yang melekat pada aktivitas itu sendiri [30]. Melalui subteori yang disebut Teori Evaluasi Kognitif (*Cognitive Evaluation Theory:*) dijelaskan bahwa peristiwa eksternal seperti imbalan atau umpan balik **mempengaruhi motivasi intrinsik berdasarkan makna fungsionalnya bagi penerima, apakah bersifat informasional yang mendukung kompetensi, atau bersifat mengontrol yang membatasi otonomi**. Peristiwa yang memfasilitasi persepsi lokus kausalitas internal akan meningkatkan motivasi intrinsik sedangkan peristiwa yang dipersepsikan mengontrol perilaku akan menggeser lokus kausalitas menjadi eksternal dan menurunkan motivasi intrinsik.

Teori ini juga dijelaskan motivasi ekstrinsik melalui Teori Integrasi Organismik (*Organismic Integration Theory*) yang menggambarkan proses internalisasi regulasi eksternal ke dalam struktur diri individu [30]. Proses internalisasi ini berlangsung dalam sebuah rangkaian mulai dari regulasi eksternal yang sepenuhnya dikendalikan oleh kontingensi luar menuju introjeksi identifikasi dan akhirnya integrasi di mana nilai eksternal berasimilasi sepenuhnya dengan diri sendiri. **Semakin terintegrasi suatu regulasi eksternal maka perilaku tersebut menjadi semakin otonom (self-determined) dan pada akhirnya berkontribusi pada kesehatan psikologis serta kualitas kinerja yang lebih baik dibandingkan regulasi yang hanya bersifat patuh atau terkontrol.**

2.4.1.4 *Positive Psychology* (Appreciation)

Psikologi positif memiliki misi fundamental untuk memahami serta membina faktor faktor yang memungkinkan individu komunitas dan masyarakat untuk berkembang atau *flourish* secara optimal [31]. Dalam perspektif ini emosi positif tidak sekadar dipandang sebagai penanda kesejahteraan subjektif semata melainkan juga berperan aktif dalam menghasilkan pertumbuhan psikologis yang berkelanjutan. Berbeda dengan **pendekatan tradisional yang seringkali berfokus pada emosi negatif** pendekatan ini menekankan bahwa **emosi positif** seperti kegembiraan ketertarikan kepuasan dan cinta memiliki nilai adaptif yang krusial karena emosi tersebut **berfungsi sebagai sarana untuk mencapai pertumbuhan psikologis** dan meningkatkan **kesejahteraan individu**, dalam hal ini mahasiswa, dalam jangka panjang [31].

Kerangka teoritis utama yang menjelaskan mekanisme fungsi emosi positif ini disebut sebagai ***broaden and build theory*** yang mempostulatkan bahwa pengalaman emosi positif memperluas atau ***broaden repertoar*** pemikiran dan tindakan sesaat individu. [31] Mekanisme perluasan ini berbeda secara signifikan dari emosi negatif yang cenderung menyempitkan fokus pikiran dan tindakan pada perilaku spesifik untuk kelangsungan hi-

dup segera dalam situasi mengancam. Sebaliknya emosi positif memperluas cakupan kognisi dan perhatian yang memungkinkan munculnya fleksibilitas kognitif kreativitas serta keinginan untuk mengeksplorasi dan mengintegrasikan informasi baru yang pada akhirnya memperkaya repertoar pemikiran dan tindakan individu pada saat tersebut.

Konsekuensi kumulatif dari perluasan pola pikir tersebut adalah terbangunnya atau *build* berbagai sumber daya personal yang bersifat tahan lama atau *enduring personal resources* yang mencakup sumber daya fisik intelektual sosial dan psikologis. Sumber daya yang terakumulasi ini berfungsi sebagai cadangan berharga yang dapat dimanfaatkan kembali di masa depan untuk mengelola ancaman serta memfasilitasi ketahanan psikologis atau resiliensi melalui efek pemulihan atau *undoing effect* terhadap dampak fisiologis emosi negatif. Dengan demikian kapasitas untuk mengalami emosi positif merupakan kekuatan fundamental yang dapat memicu spiral ke atas atau *upward spirals* menuju peningkatan kesejahteraan emosional dan transformasi diri yang lebih baik [31].

2.4.2 *Large Language Models* (LLM)

Pada bagian ini dibahas konsep dasar mengenai *Large Language Models* (LLM) yang menjadi komponen inti dari *Feedback Agent*. Pembahasan mencakup arsitektur LLM, karakteristik model Llama, serta implikasi perbedaan ukuran model (8B vs 70B) terhadap kapasitas penalaran dan performa dalam menghasilkan umpan balik berkualitas tinggi.

2.4.2.1 Dasar Teori LLM

Kecerdasan Buatan adalah cabang ilmu komputer yang berfokus pada pembuatan sistem yang mampu meniru kecerdasan manusia dalam pemecahan masalah dan pengambilan keputusan. Pendekatan awal AI banyak bersifat *symbolic* (*rule-based*), namun berkembang menuju metode *data-driven* seperti *machine learning*. Sejak awal, para peneliti AI telah membayangkan tercapainya *general problem solver* yang mampu menyelesaikan berbagai tugas secara cerdas [1]. Kemajuan mutakhir di bidang *natural language processing* menunjukkan tanda-tanda menuju visi tersebut, contohnya model bahasa besar (*large language model*) seperti *ChatGPT* yang dilatih dengan miliaran parameter pada seluruh teks internet mampu melakukan beragam tugas tekstual secara *out-of-the-box* dan menunjukkan kemampuan *emergent* mendekati kecerdasan umum manusia [1]. Perkembangan ini menandai lompatan signifikan dalam upaya mewujudkan AI yang lebih cerdas dan serbaguna.

Large Language Models (LLM) merepresentasikan terobosan signifikan dalam ranah Pemrosesan Bahasa Alami (*Natural Language Processing* atau NLP). Teknologi ini mendayagunakan jaringan saraf tiruan dengan miliaran parameter untuk mengidentifikasi pola bahasa dan memproduksi teks yang kualitasnya setara dengan komunikasi

manusia. Melalui pelatihan pada himpunan data berskala masif dengan metode *self-supervised learning*, LLM mampu melaksanakan berbagai tugas NLP yang kompleks, meliputi pembangkitan teks, penerjemahan, peringkasan, sistem tanya-jawab, hingga analisis sentimen. Kehadiran teknologi ini telah mengubah lanskap NLP secara mendasar dengan mengatasi keterbatasan metode konvensional berbasis aturan dan statistik, seperti *n-grams*, yang sebelumnya kurang efektif dalam menangkap ketergantungan jangka panjang maupun nuansa kontekstual dalam teks [35].

Sejarah perkembangan LLM dicirikan oleh kemajuan bertahap dalam arsitektur jaringan saraf. Fase awal dimulai dengan penggunaan *Recurrent Neural Networks* (RNNs) di mana digunakan untuk pengolahan data sekuensial, meskipun model ini sering terkendala oleh isu *vanishing gradients* yang menghambat proses pembelajaran jangka panjang. Kemajuan berikutnya dicapai melalui penerapan *word embeddings* dan representasi vektor yang meningkatkan pemahaman semantik dengan memetakan relasi kata dalam ruang berdimensi tinggi. Namun, lompatan teknologi terbesar terjadi melalui introduksi arsitektur *transformer* yang menerapkan mekanisme *self-attention*. Mekanisme ini memungkinkan model untuk memahami hubungan kontekstual secara jauh lebih akurat, sekaligus menjadi landasan bagi pengembangan model-model canggih kontemporer seperti GPT dan BERT.

Penggunaan LLM telah meluas ke berbagai sektor industri, termasuk di antaranya di bidang pendidikan. Dalam konteks pendidikan, LLM dimanfaatkan untuk menghasilkan materi pembelajaran yang dipersonalisasi, memberikan umpan balik otomatis kepada siswa, serta mendukung sistem tanya jawab interaktif.

BAB III
METODE PENELITIAN

Bab ini menjelaskan metode atau cara yang digunakan dalam penelitian ini untuk mencapai maksud dan tujuan seperti yang tertulis dalam Subbab 1.3 yaitu:

1. Merancang metode yang sesuai untuk meningkatkan performa LLM dalam memberikan umpan balik terhadap proses SRL mahasiswa jika dibandingkan dengan ahli manusia.
2. Mengevaluasi kualitas umpan balik yang dihasilkan LLM dari berbagai skema metode yang dirancang terhadap proses SRL mahasiswa baik secara kuantitatif maupun kualitatif dengan umpan balik ahli manusia.
3. Melihat pengaruh besar ukuran model LLM terhadap kualitas umpan balik yang dihasilkan.

3.1 Alat dan Bahan Tugas Akhir

3.1.1 Alat Tugas Akhir

Alat-alat yang digunakan pada tugas akhir ini berupa perangkat keras maupun perangkat lunak sebagai sarana pendukung. Alat-alat perangkat keras yang digunakan merupakan milik pribadi. Fungsi dari perangkat keras dan lunak ini adalah untuk penulisan kode, pengembangan chatbot, melakukan pengujian baik secara kuantitatif maupun kualitatif, serta untuk pengumpulan dataset.Berikut adalah daftar alat-alat tugas akhir yang digunakan pada penelitian ini.

1. **Macbook Pro 14 inci.** Spesifikasi utama:
   - Prosesor: M1 Pro 10 Core @ 3,2 GHz
   - GPU internal: M1 Pro GPU 16 Core
   - Neural Engine: 32 Core
   - RAM: *Unified Memory* 16 GB 200GB/s *memory bandwidth*
   - Penyimpanan: *Solid State Drive* internal 1 TB (MacOS 26 Tahoe)
2. **Visual Studio Code** versi 1.105.1 sebagai *integrated development environment* (IDE) untuk menulis dan mengelola kode.
3. **Python 3.12.7** sebagai bahasa pemrograman utama untuk mengembangkan *chatbot* berbasis LLM, dengan beberapa pustaka inti:
   - `LangChain`: kerangka kerja aplikasi LLM yang menyediakan komponen modular untuk orkestrasi *prompt*, alur data, dan integrasi sumber data eksternal.

- `LangGraph`: ekstensi `LangChain` untuk membangun alur aplikasi yang lebih kompleks dalam bentuk graf.
- `bert_score`: sebagai salah satu kerangka kerja evaluasi kuantitatif untuk mengukur kesamaan semantik (Precision, Recall, F1) antara respons yang dihasilkan model dengan data ground truth.
- `BARTScorer` sebagai kerangka kerja evaluasi kuantitatif alternatif (menggunakan mBART) untuk menghitung skor F1, Precision, dan Recall berdasarkan log-probability token.
- `Pandas` sebagai library yang digunakan untuk menyusun, mengagregasi, dan menampilkan tabel ringkasan hasil evaluasi kuantitatif (BARTScore dan GPT-4o Judge) agar mudah dibaca.

4. **Google Docs**: sebagai alat untuk menulis kuesioner, dan melakukan revisi serta dokumentasi hasil diskusi.
5. **Google Sheets**: digunakan untuk memvisualisasikan dan mengelola data Kanban sehingga dapat dibagikan kepada para ahli yang akan mengisi dan menilai data, memudahkan kolaborasi, pengisian respons, serta pengumpulan metrik penilaian.
6. **Anaconda** untuk manajemen pustaka Python dan *virtual environment* selama proses pengembangan.
7. **ChatGPT 5 Thinking** untuk membuat data sintetis papan Kanban yang akan digunakan dalam penelitian ini.
8. **Groq API** untuk melakukan inferensi LLM (pemanggilan model) yang digunakan dalam penelitian ini.

3.1.2 Bahan Tugas Akhir

Bahan yang digunakan dalam proses pembuatan chatbot dan untuk uji coba pada penelitian ini adalah sebagai berikut:

1. LLM dengan model **Llama 3.1 8B Instant**, dengan parameter delapan miliar, yang diperoleh dari platform Groq API yang berfungsi sebagai agen pemberi umpan balik (*Feedback Agent*).
2. LLM dengan model **Llama 3.3 70B Versatile**, dengan parameter lebih besar yaitu 70 miliar yang juga diperoleh dari platform Groq API yang berfungsi sebagai agen pemberi umpan balik (*Feedback Agent*).
3. **GPT-4o** sebagai model yang digunakan untuk evaluasi kuantitatif *LLM-as-a-judge*
4. **Dataset sintetis** dengan format data `.json` yang di-*generate* dari ChatGPT, berisikan pasangan data pembelajaran mahasiswa berbasis papan Kanban.

5. **Data umpan balik** berupa komentar dari ahli pembelajaran yang berisikan *feedback*, *motivation*, dan *appreciation*.
6. **Data hasil kuesioner** dari ahli psikologi pendidikan yang berisikan penilaian dan komentar terhadap hasil umpan balik LLM dalam menilai performa belajar mahasiswa.

3.2 Metode yang Digunakan

3.2.1 Metode Pembuatan Data Sintetis

Penelitian ini memerlukan sebuah *dataset* khusus yang mampu merefleksikan proses belajar mahasiswa dalam platform *learning analytics* berbasis Kanban, sekaligus memiliki *ground truth* pedagogis berkualitas tinggi. Karena *dataset* publik dengan spesifikasi tersebut tidak tersedia, dirancang sebuah metodologi komprehensif untuk pembuatan data sintetis. Proses ini bertujuan untuk menghasilkan *dataset* yang realistis dan bervariasi, yang mencakup dua komponen utama: (1) data jejak digital (*digital trace data*) mahasiswa dalam format `JSON` yang merepresentasikan aktivitas papan Kanban mereka, dan (2) data umpan balik ahli manusia yang selaras dengan setiap data jejak tersebut.

Gambaran besar dari metodologi ini dimulai dengan analisis mendalam terhadap platform pembelajaran Gamatutor [36]. Analisis ini berfokus pada dekonstruksi skema *database* MongoDB yang ada untuk mengidentifikasi atribut-atribut kunci yang merefleksikan proses SRL mahasiswa, seperti yang terlihat pada Gambar 3.7. Untuk memfokuskan proses generasi data, skema ini kemudian direduksi menjadi komponen-komponen yang paling esensial, sebagaimana dirinci dalam Gambar 3.8.

Langkah krusial berikutnya adalah mendefinisikan realisme perilaku. Untuk menangkap spektrum keterlibatan mahasiswa yang luas, dirumuskan lima kategori profil perilaku yang berbeda, mulai dari `"MALAS"` hingga `"SANGAT RAJIN"`. Setiap profil ini dioperasionalkan ke dalam serangkaian aturan studi kasus yang ketat, yang menentukan variabel terukur seperti jumlah total kartu, persentase penyelesaian *checklist*, dan aktivitas pergerakan kartu terbaru.

Untuk memastikan relevansi konten, *dataset* ini juga dibatasi pada satu domain mata kuliah tertentu, yaitu Pemrograman Berbasis Objek (OOP). Lima subtopik spesifik dari mata kuliah ini telah didefinisikan, lengkap dengan tujuan pembelajaran dan *checklist items* yang mendetail, seperti yang dipaparkan pada Tabel 3.2.

Seluruh aturan ini mencakup skema data, profil perilaku, dan batasan konten yang kemudian ditransformasi menjadi serangkaian *prompt* instruksi yang sangat t erstruktur. *Prompt* ini dirancang menggunakan teknik *prompt engineering* lanjutan, termasuk *role-play*, *Few-shot* untuk memberikan contoh format `JSON` yang diharapkan, dan *Chain-of-*

Thought (CoT) untuk memandu logika penalaran LLM.

Menggunakan *prompt* gabungan ini [Gambar 3.15], sebuah model LLM (ChatGPT 5 Thinking) ditugaskan untuk menghasilkan 40 *dataset* papan Kanban sintetis yang unik dan valid.[1] Setelah data `JSON` mentah berhasil dibuat, proses dilanjutkan ke tahap pengumpulan *ground truth*. Data `JSON` yang kompleks tersebut ditransformasi oleh LLM menjadi format tabel yang mudah dibaca manusia, dengan menghilangkan label kategori untuk memastikan penilaian yang netral.

Terakhir, data tabular ini disajikan kepada ahli manusia (dosen) melalui Google Sheets dan kuesioner Google Docs. Para ahli kemudian memberikan umpan balik pedagogis terstruktur yang mencakup tiga fungsi utama: *feedback*, *motivation*, dan *appreciation*. Kumpulan umpan balik ahli inilah yang akan menjadi *ground truth* definitif untuk melatih dan mengevaluasi agen LLM dalam penelitian ini. Sub-bab berikutnya akan menguraikan setiap langkah dalam alur kerja pembuatan *dataset* ini secara lebih rinci.

3.2.2 Metode Peningkatan Hasil Umpan Balik

Untuk menjawab rumusan masalah penelitian, bagian ini merinci metodologi yang digunakan untuk meningkatkan kualitas umpan balik pedagogis yang dihasilkan oleh LLMs agar mendekati standar ahli manusia. Alur peningkatan ini tidak bergantung pada satu teknik tunggal, melainkan merupakan pendekatan multifaset yang sistematis. Metodologi ini didasarkan pada tiga pilar utama yang akan diuraikan dalam sub-bab berikut: (1) rekayasa konteks (*context engineering*) secara mendalam, (2) integrasi metode *learning analytics* yang bersifat deskriptif, dan (3) perancangan eksperimen variasi *prompt* untuk menguji dampak setiap perlakuan.

Pilar pertama, *context engineering*, berfokus pada perancangan *prompt* dasar (*baseline*) yang kuat secara teoretis. *Prompt* ini dirancang untuk memenuhi tujuan-tujuan pedagogis spesifik, seperti memberikan bimbingan adaptif dan meningkatkan efikasi diri, yang berlandaskan pada literatur psikologi pendidikan [28–30]. Untuk memastikan struktur dan konsistensi, *prompt* ini dibagi menjadi lima komponen logis: Definisi Peran, Logika Internal, Informasi Kontekstual, *Prompt* Penilaian, dan Format Input/Output. Kualitas dari arsitektur *prompt* ini kemudian divalidasi menggunakan *theory-driven prompt manual* [17], di mana *prompt* yang digunakan dalam penelitian ini berhasil mencapai skor 13 dari 16, yang termasuk dalam kategori "Tinggi".

Pilar kedua adalah integrasi *learning analytics* untuk menyuntikkan kesadaran situasional (*situational awareness*) ke dalam model. Pendekatan ini mentransformasi data mentah `JSON` papan Kanban mahasiswa menjadi *insight* yang dapat ditindaklanjuti. Dengan menerapkan *Descriptive Analytics*, serangkaian 12 indikator kunci seperti `aging_wip_h`, `stuck`, dan `alerts` yang diekstraksi dari data papan. Indikator-

indikator ini didasarkan pada praktik manajemen Kanban yang telah mapan [37, 38] dan dikalibrasi dengan ambang batas 2 jam yang relevan dengan studi kasus penelitian. Alur proses analitik ini [Gambar 3.22] menghasilkan sebuah ringkasan analitik yang padat, yang kemudian dimasukkan ke dalam *prompt* untuk memungkinkan LLM memberikan umpan balik yang terpersonalisasi berdasarkan data.

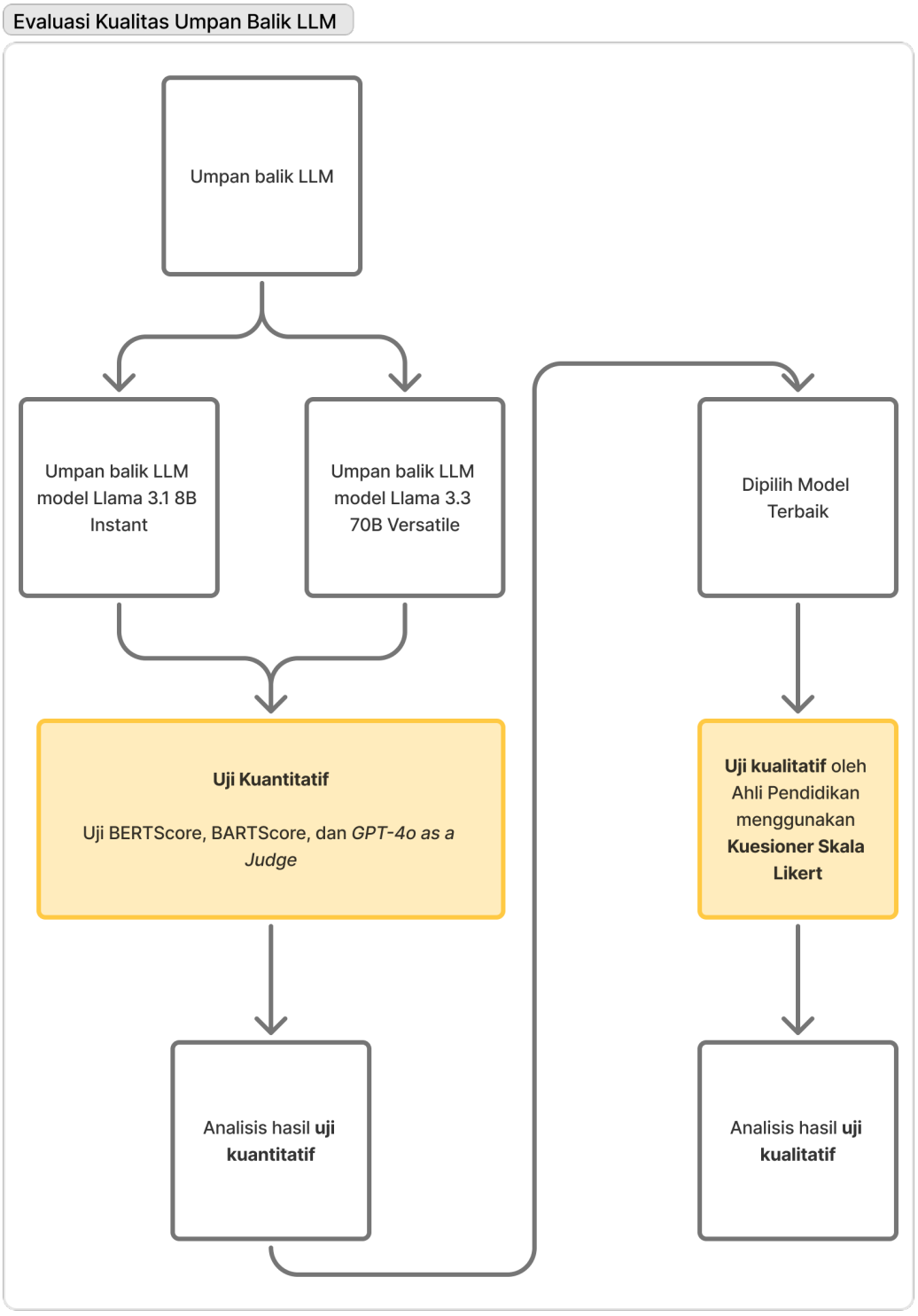
Pilar ketiga adalah pembuatan variasi *prompt* untuk mengukur efektivitas dari dua pilar pertama secara empiris. Seperti yang diilustrasikan pada Gambar 3.24, sebuah eksperimen sistematis dirancang untuk membandingkan empat kondisi *prompt* yang berbeda. Variasi ini mencakup: (1) *Baseline Prompt*, (2) *Prompt* dengan tambahan konteks "Lis Materi", (3) *Prompt* yang diperkaya dengan data *learning analytics*, dan (4) *Prompt* yang menggabungkan *learning analytics* dengan mekanisme ReAct (*Reasoning + Acting*). Mekanisme ReAct ini, yang diimplementasikan menggunakan *framework* LangGraph [Gambar 3.25], dirancang untuk memaksa model melakukan penalaran internal terhadap data analitik sebelum menghasilkan keluaran.

Seluruh eksperimen ini akan dijalankan pada dua model LLM dengan ukuran berbeda, yaitu Llama 3.1 8B Instant dan Llama 3.3 70B Versatile. Tujuan akhirnya adalah untuk menganalisis secara kuantitatif dan kualitatif dampak dari setiap strategi peningkatan—konteks, *learning analytics*, dan *ReAct* terhadap kualitas umpan balik pedagogis yang dihasilkan, sekaligus menjawab pertanyaan penelitian mengenai pengaruh ukuran model.

3.2.3 Metode Evaluasi Kualitas Umpan Balik LLM

Penelitian ini menggunakan desain eksperimen untuk melakukan perbandingan dan menilai kualitas dari beberapa variasi *context engineering* yang digabungkan dengan *learning-analytics* dalam meningkatkan mutu umpan balik LLM berbahasa Indonesia pada skenario SRL berbasis papan Kanban. Setiap variasi *context engineering* diimplementasikan secara terpisah dan dipadukan dengan dua ukuran model (Llama 3.1 8B Instant dan Llama 3.3 70B Versatile melalui Groq API), menghasilkan beberapa kombinasi umpan balik yang diuji secara kuantitatif mapun kualitatif. Pengujian kuantitatif berfokus pada kesamaan semantik (Subbab [2.1.7]) menggunakan metrik berbasis *embedding* kontekstual, yaitu BERTScore [22] dan BARTScore [23], untuk mengukur keselarasan antara respons LLM dengan data umpan balik ahli manusia *(ground truth)* [32, 33]. Selain itu, evaluasi relevansi juga dilakukan menggunakan pendekatan *LLM-as-a-Judge* yang terstruktur [24, 25]. Di sisi lain, pengujian kualitatif yang merupakan standar emas Subbab 2.1.8 yang melibatkan peninjauan langsung oleh pakar psikologi pendidikan [33]. Pakar ini menilai keluaran LLM menggunakan rubrik penilaian komprehensif yang didasarkan pada kerangka teoretis mapan seperti model Hattie dan Timperley (2007) [29], Shute (2008) [28], dan *Self-Determination Theory* [30]. Penilaian ini mengukur tiga di-

mensi utama yakni *feedback*, *motivation support*, dan *appreciation support* menggunakan skala Likert (1-5) untuk memastikan kualitas pedagogis dan psikologis dari umpan balik tersebut (Subbab[2.1.8]). Berikut adalah diagram alur metode evaluasinya.



Gambar 3.3. Diagram Alir Metode Evaluasi.

Terlihat dari Gambar 3.3, alur evaluasi penelitian ini dimulai dengan generasi umpan balik LLM yang dipadukan dengan berbagai skema *context engineering*. Umpan balik ini dihasilkan oleh dua model berbeda untuk perbandingan, yaitu Llama 3.1 8B *Instant* dan Llama 3.3 70B *Versatile*.

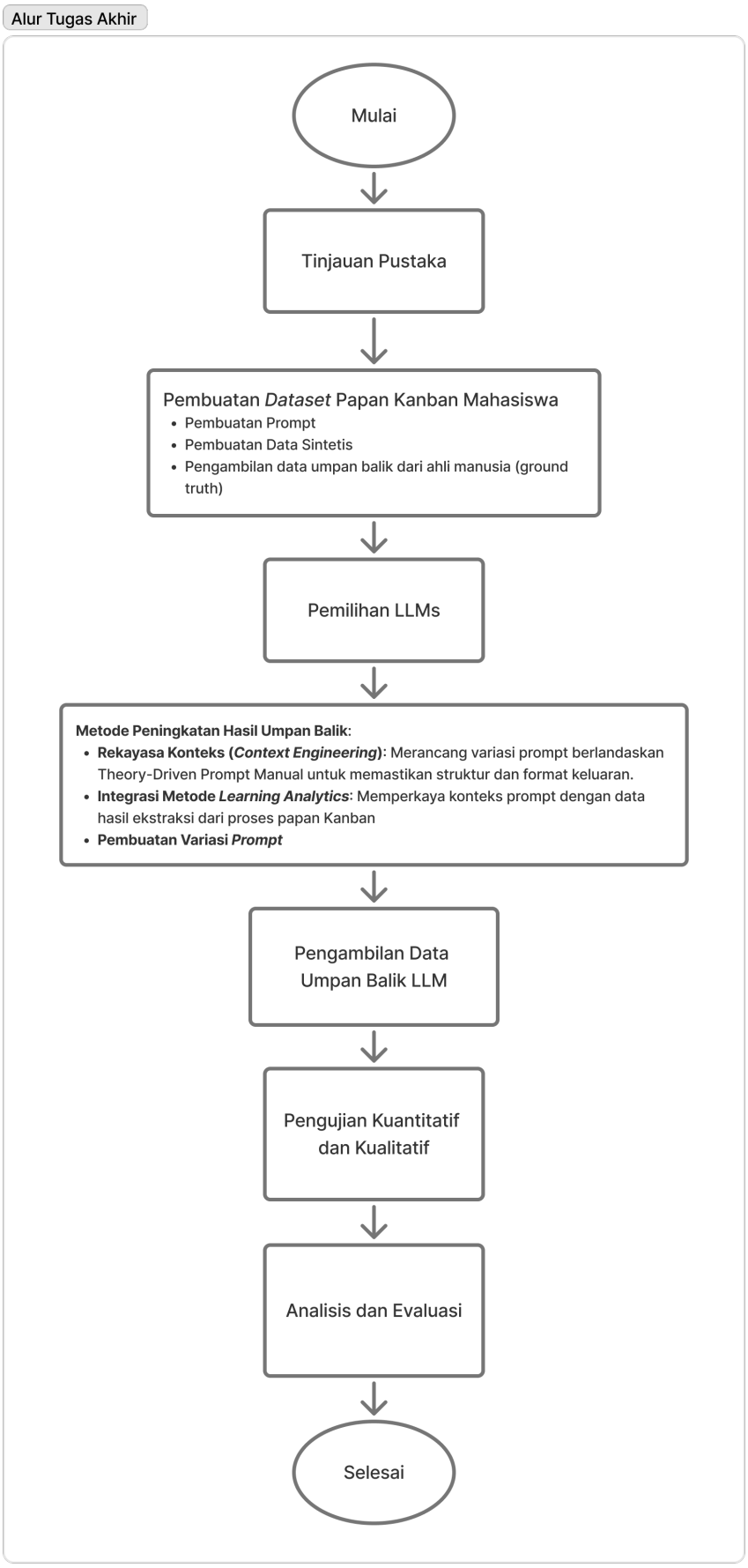
Luaran dari model-model ini pertama-tama melalui tahap Uji Kuantitatif. Sejalan dengan metodologi yang telah dijelaskan, tahap ini berfokus pada pengukuran kesamaan semantik dan relevansi Subbab 2.1.7 menggunakan tiga metrik utama: Uji BERTScore, BARTScore, dan *LLM-as-a-Judge* [22, 23, 24].

Luaran dari pengujian ini kemudian menjalani analisis hasil Uji Kuantitatif. Berdasarkan analisis ini, model terbaik dipilih—yaitu, kombinasi model dan teknik *context engineering* yang menghasilkan skor keselarasan tertinggi terhadap *ground truth* ahli.

Selanjutnya, umpan balik dari model terbaik tersebut dilanjutkan ke tahap evaluasi kedua, yaitu Uji Kualitatif. Tahap ini merupakan "standar emas" Subbab 2.1.8 yang melibatkan peninjauan langsung oleh Ahli Pendidikan. Pakar ini menilai kualitas pedagogis dan psikologis dari umpan balik model terbaik menggunakan Kuesioner Skala Likert yang dirancang khusus berdasarkan kerangka teoretis [26, 27, 28]. Langkah terakhir adalah Analisis hasil uji kualitatif untuk mendapatkan kesimpulan akhir mengenai efektivitas umpan balik LLM dalam konteks SRL.

3.3 Alur Tugas Akhir

Dalam penelitian ini, terdapat sebuah alur kerja yang menjabarkan semua tahapan dan proses yang terlibat. Secara keseluruhan, proses penelitian terdiri dari empat tahap utama: tahap studi pendahuluan dan persiapan data, yang mencakup Tinjauan Pustaka dan Pembuatan *Dataset* Papan Kanban Mahasiswa (termasuk pengumpulan *ground truth* ahli); tahap desain eksperimen, yang meliputi Pemilihan Model LLMs dan perancangan Metode Peningkatan Hasil Umpan Balik (menggunakan Rekayasa Konteks, Integrasi *Learning Analytics*, dan Pembuatan Variasi *Prompt*); tahap implementasi, di mana dilakukan Pengambilan Data Umpan Balik LLM menggunakan metode yang telah dirancang; serta tahap evaluasi akhir, yang terdiri dari Pengujian Kuantitatif dan Kualitatif, diikuti oleh Analisis dan Evaluasi. Alur kerja penelitian secara keseluruhan diilustrasikan pada Gambar 3.4.



Gambar 3.4. Flowchart Alur Penelitian Tugas Akhir.

3.3.1 Tinjauan Pustaka

Bab ini menyajikan tinjauan pustaka mendalam yang relevan dengan topik penelitian. Tinjauan ini berfokus pada penelitian-penelitian sebelumnya mengenai peningkatan kualitas umpan balik LLM dalam konteks pendidikan, dengan menyelidiki dua arus utama yang teridentifikasi dari karya-karya kunci [Tabel 2.1]. Arus pertama adalah penyelarasan model (*model alignment*) melalui *fine-tuning*, seperti *Direct Preference Optimization* (DPO) [13] dan *Reinforcement Learning from Human Feedback* (RLHF) [16]. Arus kedua adalah penyelarasan konteks (*context alignment*), yang mencakup rekayasa *prompt* (*prompt engineering*) berbasis teori [15] dan rekayasa konteks yang digabungkan dengan *learning analytics* [17]. Tinjauan ini juga menganalisis pemanfaatan dasbor orkestrasi kelas berbasis Kanban [18] sebagai sumber data proses, bersama dengan metode evaluasi yang digunakan untuk menilai kinerja umpan balik, baik secara kuantitatif (misalnya, BERTScore [22], BARTScore [23], dan *LLM-as-a-Judge* [24]) maupun secara kualitatif oleh ahli manusia sebagai standar emas [26, 27, 28].

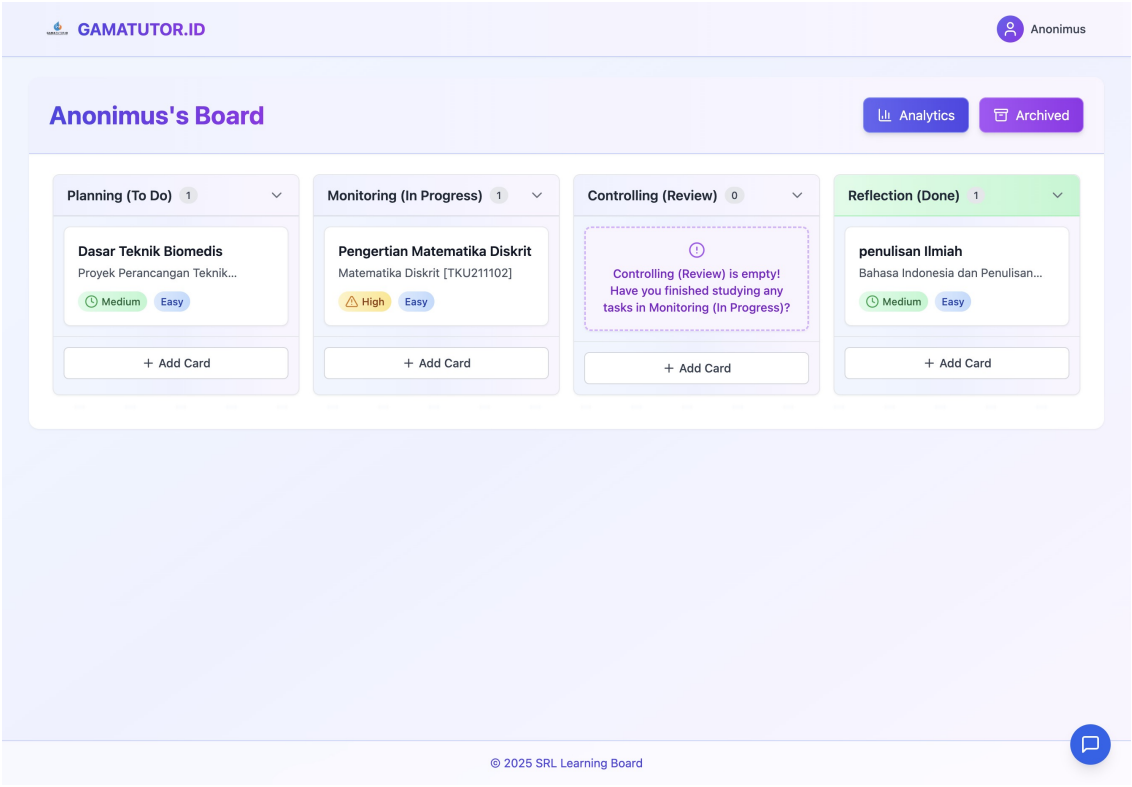
Selanjutnya, analisis terhadap metodologi yang diterapkan dan temuan utama dalam studi-studi ini dilakukan. Tujuan utama dari analisis ini adalah untuk menilai secara kritis keunggulan dan keterbatasan dari setiap pendekatan. Evaluasi ini mengidentifikasi keunggulan (seperti efisiensi biaya dan kelayakan implementasi dari rekayasa konteks [15]) dan kelemahan (seperti tingginya kebutuhan sumber daya komputasi dan data untuk *fine-tuning* [13, 16]). Temuan ini menjadi pertimbangan dalam pemilihan dan adaptasi metodologi yang paling tepat untuk menjawab masalah penelitian dalam karya ini: memanfaatkan pendekatan *context engineering* yang ringan dan murah [15] namun diperkaya dengan data *learning analytics* kaya sinyal yang diekstraksi dari papan Kanban [18] untuk mengatasi celah penelitian yang teridentifikasi.

3.3.2 Pembuatan Dataset Papan Kanban Mahasiswa

Pada bagian ini, dilakukan proses pembuatan *Dataset* Papan Kanban Mahasiswa. Tahap ini mencakup beberapa langkah krusial, yaitu Pembuatan *Prompt* , Pembuatan Data Sintetis yang akan menjadi data papan Kanban , serta Pengambilan data umpan balik pedagogis dari ahli manusia yang akan digunakan sebagai *ground truth* untuk perbandingan.

3.3.2.1 Pembuatan *Prompt*

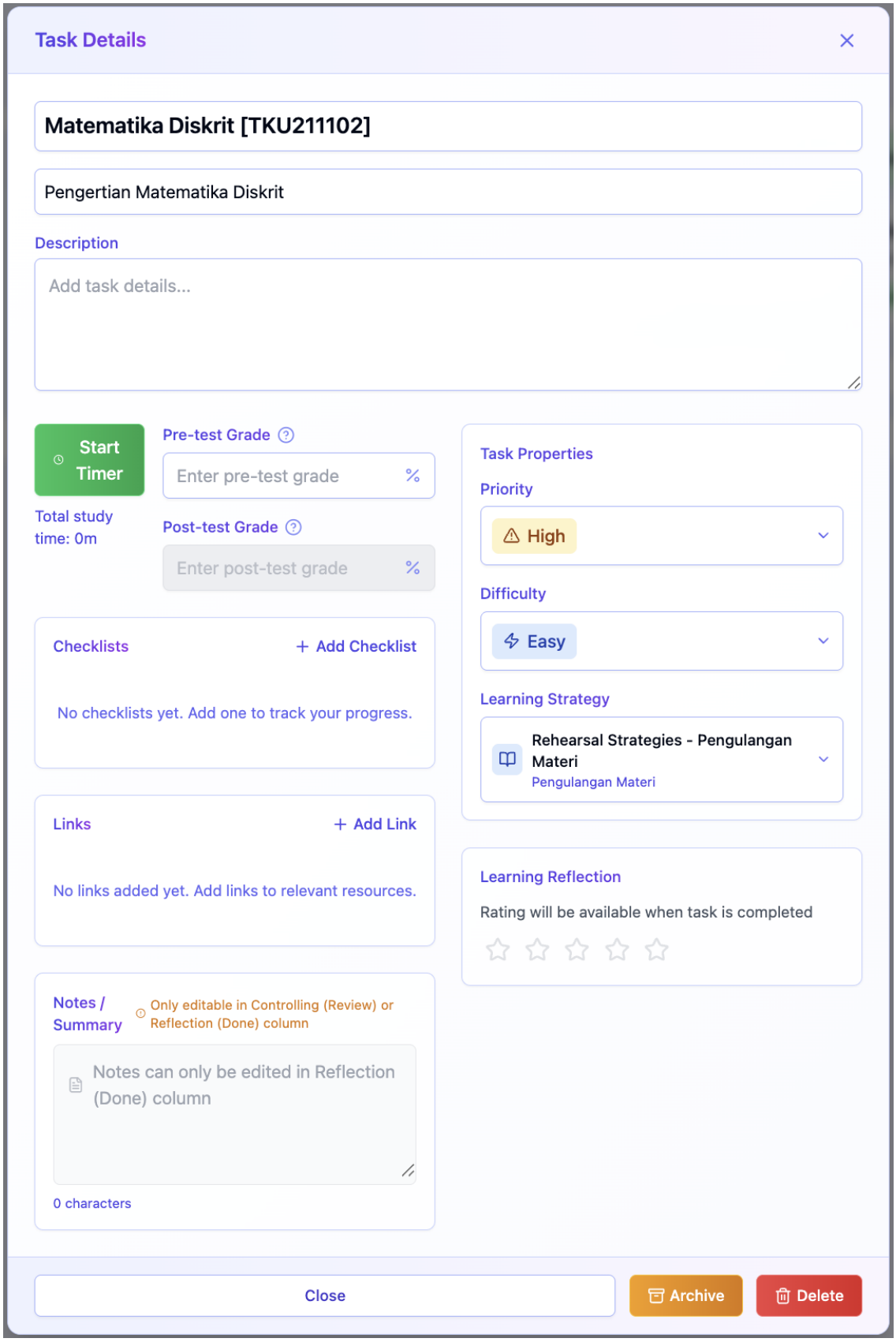
Proses pembuatan *prompt* dimulai dengan analisis mendalam terhadap struktur *database* asli dari platform pembelajaran berbasis papan Kanban. Platform yang digunakan dan dikembangkan bernama Gamatutor yang dikembangkan oleh M. I. Azmi [36]. Berikut adalah tampilan dari platform pembelajaran tersebut yang ditampilkan pada Gambar 3.5.



Gambar 3.5. Tampilan Utama Platform Gamatutor.

Secara umum, Gamatutor adalah platform *learning analytics* berbasis metode *Kanban* yang dirancang untuk mendukung SRL mahasiswa. Platform ini memfasilitasi pengelolaan tugas belajar melalui sistem papan visual dengan empat tahapan utama yaitu *Planning (To Do)*, *Monitoring (In Progress)*, *Controlling (Review)*, dan *Reflection (Done)*. Setiap tugas direpresentasikan dalam bentuk *kartu* yang memuat informasi lengkap seperti judul, mata kuliah, *priority*, *difficulty*, *learning strategy*, nilai *pre-test/post-test*, dan *checklist* progres. Fitur utamanya meliputi pembuatan dan pengelolaan kartu tugas, pemantauan *progress* melalui *learning analytics*, serta manajemen profil pengguna.

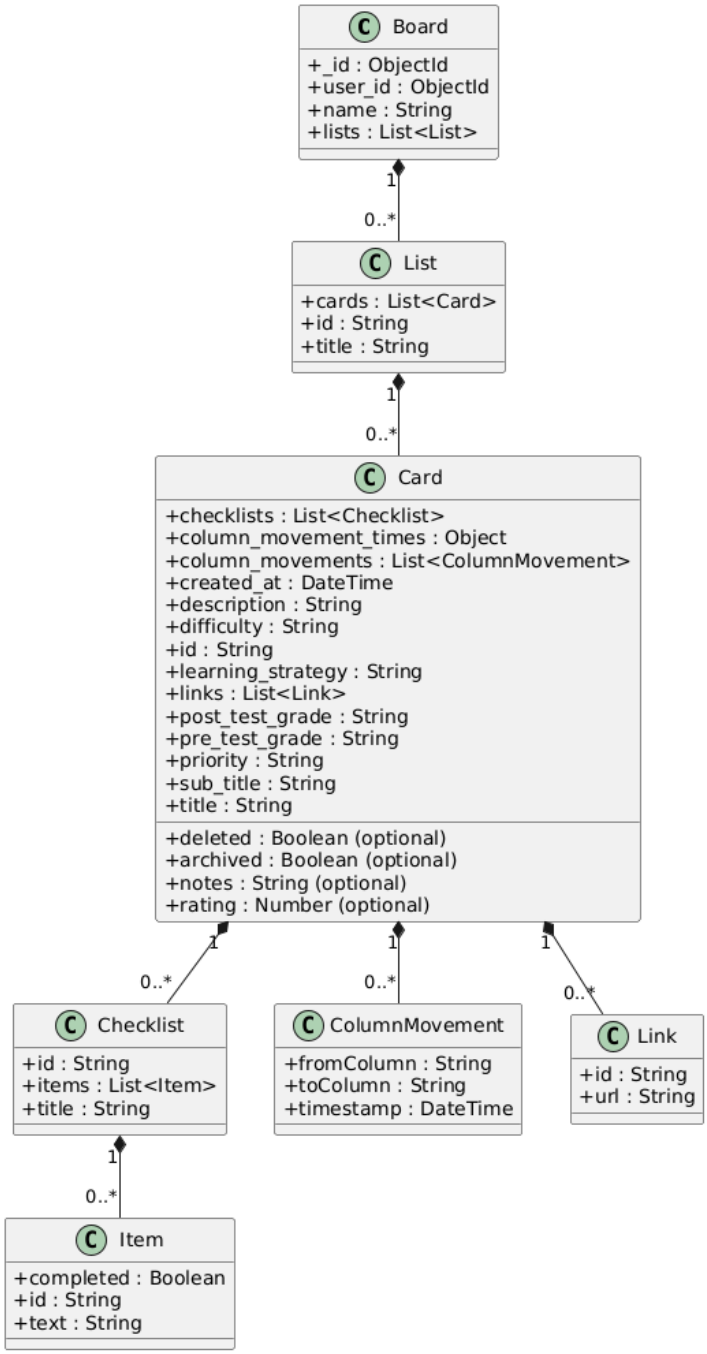
Untuk satu kartunya merepresentasikan satu subtopik dari salah satu mata kuliah. Jadi mahasiswa bisa membuat banyak kartu untuk mata kuliah yang sama. Untuk mengakses detail informasi-informasi tersebut, pengguna (mahasiswa) dapat melihatnya pada tiap-tiap kartu dengan menekan kartunya. Tampilan dari informasinya dapat dilihat pada Gambar 3.6



Gambar 3.6. Tampilan Informasi Tiap Kartu Gamatutor.

Semua detail informasi dan fitur-fitur tersebut dibuat untuk membantu mahasiswa dalam menerapkan SRL (seperti: *Start Timer*, *Checklists*, dan "Add Link") serta secara bersamaan memberikan data proses yang kaya untuk *learning analytics*. Data ini mencakup *Total study time* (yang dihasilkan dari fitur *Start Timer*), progres penyelesaian tugas melalui *Checklists*, hingga data metakognitif dan performa seperti *Priority, Difficulty, Learning Strategy* yang dipilih, serta nilai *Pre-test Grade* dan *Post-test Grade*. Kumpulan data proses inilah yang dapat menjadi insight bagi agen pedagogis (guru ataupun *AI chatbot*) untuk memantau, mendiagnosis hambatan, dan memberikan bimbingan yang terpersonalisasi sesuai dengan kondisi aktual mahasiswa tersebut. Gamatutor juga memiliki *Learning Assistant* dengan berbasiskan chatbot LLM yang rencananya akan diterapkan secara bertahap. Penelitian ini adalah bagian dari pengembangan *Learning Asistant* tersebut.

Semua informasi mengenai pembelajaran mahasiswa dan kartu-kartu Kanban-nya secara rapi disimpan ke dalam *database* NoSQL yaitu MongoDB. Data tersebut disimpan dengan mengikuti format `JSON`. Berikut adalah skema penyimpanan data kartu Kanban dalam MongoDB.



Gambar 3.7. Skema Penyimpanan Data Kartu Pembelajaran Kanban Mahasiswa.

Dapat dilihat pada Gambar 3.7 terdapat banyak atribut, khususnya pada entitas `Card`, yang menyimpan data-data kunci untuk merefleksikan proses dan hasil belajar mahasiswa. Satu `Board`, hanya bisa dimiliki oleh satu mahasiswa saja di mana `Board` sendiri adalah representasi dari keseluruhan papan pembelajaran berbasis Kanban.

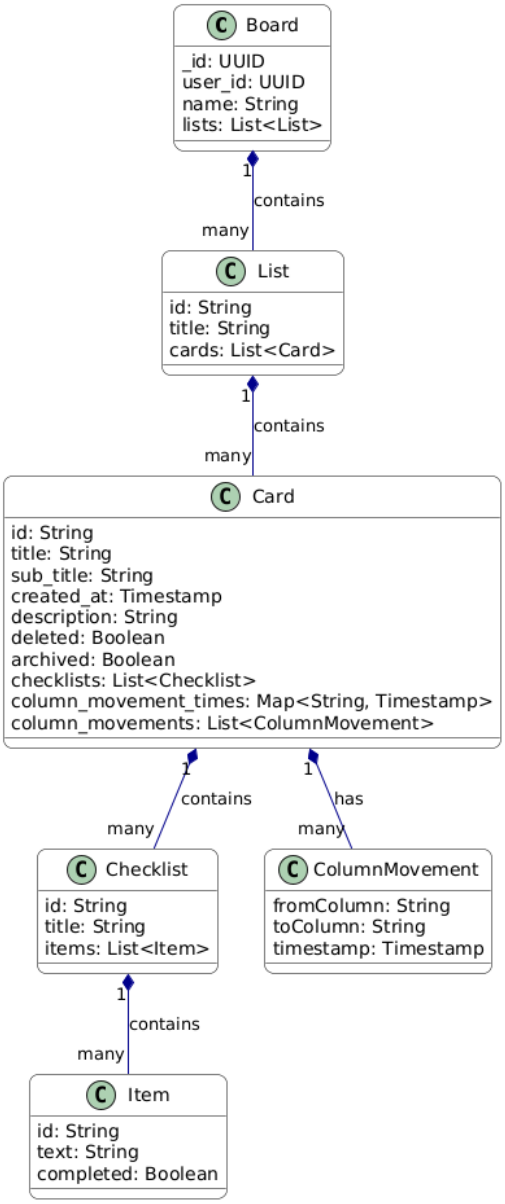
Beberapa atribut yang paling penting untuk mengevaluasi kemampuan belajar ini meliputi `pre_test_grade` dan `post_test_grade`. Kedua atribut ini secara langsung mengukur pengetahuan mahasiswa sebelum dan sesudah mengerjakan tugas, se-

hingga selisih antara keduanya dapat menjadi indikator utama peningkatan pemahaman.

Selain itu, atribut `difficulty` mengkategorikan tingkat kesulitan tugas, yang memungkinkan analisis performa mahasiswa terhadap tantangan yang diberikan. Atribut `column_movement_times` juga sangat kaya data dengan `timestamp` di dalamnya, data ini dapat digunakan untuk menghitung total durasi pengerjaan, waktu yang dihabiskan dalam fase belajar aktif (di kolom "In Progress"), atau bahkan mendeteksi adanya revisi. Terdapat juga atribut `notes` yang berisikan catatan dari mahasiswa ketika kartu sudah berada di tahap *Controlling (Review)*. Isi catatan ini bisa menjadi gambaran penting tentang rangkuman pembelajaran yang dilakukan mahasiswa terhadap subtopik tersebut yang menjadikannya data yang cukup penting.

Atribut lain yang relevan adalah `checklists` untuk melihat progres granular pengerjaan sub-tugas, `learning_strategy` untuk memahami pendekatan belajar yang dipilih mahasiswa, dan `rating` untuk memberikan penilaian terhadap pengalaman belajar terhadap subtopik yang dipelajari. Kombinasi dari atribut-atribut inilah yang memberikan gambaran komprehensif mengenai pola dan kemampuan belajar mahasiswa.

Dari skema MongoDB diatas, dilakukan reduksi atribut untuk memfokuskan pada data yang paling esensial bagi penelitian, seperti `created_at`, `checklists`, dan `column_movements`. Selain itu, dilakukan reduksi ini untuk mempermudah pembuatan data sintetis. Berikut adalah hasil skema database MongoDB setelah dilakukan reduksi data yang ada pada Gambar 3.8.



Gambar 3.8. Skema Penyimpanan Data Kartu Pembelajaran Kanban Mahasiswa Setelah Reduksi

Untuk mencapai tingkat representasi yang valid, data tersebut harus mampu menangkap berbagai pola perilaku dan tingkat keterlibatan mahasiswa yang berbeda-beda, mulai dari yang paling tidak aktif hingga yang sangat proaktif. Sebagai strategi utama untuk mengimplementasikan keragaman ini, sebuah studi kasus telah dirumuskan yang berfokus pada pengkategorian mahasiswa ke dalam profil-profil perilaku yang spesifik. Melalui studi kasus ini, telah diidentifikasi dan ditetapkan lima kategori profil yang berbeda, di mana setiap kategori mewakili satu titik pada spektrum tingkat kerajinan mahasiswa. Kelima kategori yang komprehensif ini yang diberi label `"MALAS"`, `"Sedikit Malas"`, `"Cukup"`, `"Rajin"`, dan `"Sangat Rajin"`, akan digunakan sebagai dasar untuk menghasilkan set data yang bervariasi.

Untuk menyederhanakan penelitian ini, penggunaan asumsi akan digunakan dalam memberikan label kategori tingkat performa dan perilaku mahasiswa. Variabel atribut yang digunakan antara lain adalah jumlah total `card` yang dibuat oleh mahasiswa, tingkat penyelesaian tugas (`checklist`), aktivitas terkini dari kartu tersebut pada `column_movements`, dan letak keberadaan `card` tersebut. Hal ini diharapkan dapat menjadi proksi yang *memadai dan terukur* untuk membedakan secara jelas antara kelima profil perilaku dan performa tersebut, sehingga penyederhanaan ini tetap dapat menghasilkan data yang representatif.

Aturan-aturan tersebut dirumuskan dan diimplementasikan pada studi kasus seperti jumlah total kartu, persentase penyelesaian *checklist*, dan stempel waktu pergerakan kartu yang secara implementasinya dapat dilihat pada Tabel 3.1.

Tabel 3.1. Tabel Aturan Studi Kasus Perilaku Mahasiswa

Kategori	**Studi Kasus**	**Jumlah Kartu**	**Penyelesaian Checklist (per kartu)**	**Aturan Aktivitas & Posisi Kartu**
Malas	MHS_MALAS_SK1	1 - 2 kartu	0 - 2 item selesai (dari 3)	**Tidak ada aktivitas** (2 jam terakhir). Kartu cenderung di list1 atau list2.
Sedikit Malas	MHS_SEDMALAS_SK1	3 - 4 kartu	0 - 2 item selesai (dari 3)	**Tidak ada aktivitas** (2 jam terakhir).
	MHS_SEDMALAS_SK2	3 - 4 kartu	1 - 2 item selesai (dari 3)	**Tidak ada aktivitas** (2 jam terakhir).
Cukup	MHS_CUKUP_SK1	4 - 5 kartu	1 - 2 item selesai (dari 3)	**Ada aktivitas** (2 jam terakhir) untuk min. 3 kartu.
	MHS_CUKUP_SK2	3 - 4 kartu	Semua 3 item selesai	**Ada aktivitas** (2 jam terakhir) untuk min. 3 kartu.
Rajin	MHS_RAJIN_SK1	Tepat 5 kartu	2 - 3 item selesai (dari 3)	Hanya 1 kartu tersisa di list1 ("Planning (To Do)").
	MHS_RAJIN_SK2	4 - 5 kartu	Semua 3 item selesai	Hanya 1 kartu tersisa di list1 ("Planning (To Do)").
Sangat Rajin	MHS_SANGRAJIN_SK1	Tepat 5 kartu	Semua 3 item selesai	Semua 5 kartu berada di list4 ("Reflection (Done)").

Kategori-kategori ini kemudian dijadikan dasar untuk membuat *prompt* generator data sintetis pembelajaran mahasiswa berbasis Kanban. Kumpulan aturan yang telah dirangkum dalam Tabel 3.1 tersebut kemudian ditransformasi menjadi sebuah *prompt* instruksi yang terstruktur. Proses transformasi ini melibatkan penerjemahan setiap studi kasus—mulai dari MHS_MALAS_SK1 hingga MHS_SANGRAJIN_SK1, dari format

tabel tabular menjadi format teks naratif. Tujuan utamanya adalah untuk menyajikan aturan-aturan tersebut dalam format bahasa alami (*natural language*) yang dapat dipahami dan dieksekusi secara presisi oleh model LLM pembuat data tersebut. *Prompt* ini secara eksplisit merinci batasan untuk setiap variabel, termasuk jumlah kartu yang harus dibuat, syarat penyelesaian *checklist*, dan aturan mengenai aktivitas pergerakan kartu. Terdapat sedikit penambahan aturan pada pembuatan *prompt* yaitu adanya waktu kapan dibuatnya kartu subtopik tersebut. Contoh *prompt* lengkap yang telah diformulasikan untuk agen generator tersebut dapat dilihat secara rinci pada potongan snipet pada Gambar 3.9.

Gambar 3.9. Contoh *Prompt* untuk Studi Kasus Data Sintetis Mahasiswa.

```
"""

Kategori: Malas

* Aturan untuk Studi Kasus 1 (MHS_MALAS_SK1):

* Buat 1 papan JSON dengan 1 atau 2 kartu.

* Setiap kartu memiliki 3 checklist, dengan total 0 hingga 2 item yang
    dicentang (completed: true) per kartu.

* Tidak ada kartu yang berpindah tempat dalam kurun waktu 2 jam
   terakhir (semua timestamp created_at atau column_movements terakhir
    harus <= 2025-06-07T15:00:00Z). Kartu cenderung berada di list1
   atau list2.



Kategori: Sedikit Malas

* Aturan untuk Studi Kasus 1 (MHS_SEDMALAS_SK1):

* Buat 1 papan JSON dengan 3 atau 4 kartu.

* Setiap kartu memiliki 3 checklist, dengan total 0 hingga 2 item yang
    dicentang per kartu.

* Tidak ada kartu yang berpindah tempat dalam kurun waktu 2 jam
   terakhir.

* Aturan untuk Studi Kasus 2 (MHS_SEDMALAS_SK2):

* Buat 1 papan JSON dengan 3 atau 4 kartu.

* Setiap kartu memiliki 3 checklist, dengan total 1 hingga 2 item yang
    dicentang per kartu.

* Tidak ada kartu yang berpindah tempat dalam kurun waktu 2 jam
   terakhir.


# .... Diatas adalah potongan yang tidak lengkap dari kode yang utuh.
   ....

"""
```

Selanjutnya, untuk memperkecil topik pembelajaran yang akan dibuat pada data sintetisnya, dibatasi satu mata kuliah saja dengan 5 subtopik yang dapat diambil. Pembuatan batasan subtopik harus memiliki aturan yang perlu diberikan ke model LLM pembuat data sintetis tersebut. Hal ini agar model dapat memberikan data luaran yang baik serta variasi yang beragam dengan batasan yang jelas. Tabel 3.2 menunjukkan batasan subtopik yang akan diberikan pada data sintetis.

Tabel 3.2. Ringkasan Topik OOP, Tujuan, dan Checklist

Topik	**Tujuan**	**Checklist Items**
Topik 1: Konsep Dasar dan Pilar OOP	Memahami fondasi dari OOP dan 4 pilar utamanya.	• Menjelaskan perbedaan paradigma pemrograman prosedural dan OOP. • Mengidentifikasi serta mendeskripsikan 4 pilar OOP: enkapsulasi, abstraksi, pewarisan, polimorfisme. • Menerapkan konsep dasar objek dan kelas dalam program sederhana.
Topik 2: Perancangan Kelas dan Interaksi Objek	Mampu merancang struktur kelas dan interaksinya secara logis.	• Mendesain *class diagram* berdasarkan kebutuhan dari studi kasus. • Menentukan relasi antar objek: asosiasi, agregasi, komposisi. • Mengimplementasikan rancangan interaksi objek dalam bentuk kode.
Topik 3: Enkapsulasi, Inheritance, dan Polimorfisme	Menguasai teknik pewarisan dan fleksibilitas perilaku objek.	• Menerapkan enkapsulasi menggunakan modifier akses (private, public, protected). • Membuat kelas turunan dan meng-*override* metode dari *superclass*. • Menggunakan polimorfisme melalui *overloading* dan *overriding* metode.
Topik 4: Prinsip SOLID dan Adaptive Code	Menulis kode yang fleksibel, *maintainable*, dan *scalable*.	• Menjelaskan dan memberi contoh penerapan dari masing-masing prinsip SOLID. • Melakukan *refactoring* kode agar sesuai dengan prinsip SOLID. • Menerapkan prinsip desain adaptif dalam proyek mini berbasis OOP.
Topik 5: Kolaborasi Objek dan Pengorganisasian Kode	Menyusun kode OOP secara terstruktur dan profesional.	• Mengorganisasi struktur folder proyek OOP berdasarkan lapisan fungsional. • Mengimplementasikan interaksi antar objek menggunakan *interface* atau *abstract class*. • Mengembangkan aplikasi kecil dengan arsitektur OOP yang modular.

Tabel 3.2 merincikan batasan konten pembelajaran yang digunakan dalam penelitian ini. Fokus mata kuliah dikerucutkan pada Pemrograman Berbasis Objek (*Object Oriented Programming* atau OOP). Tabel tersebut menjabarkan lima subtopik utama yang akan digunakan sebagai materi kartu tugas, mulai dari "Konsep Dasar dan Pilar OOP" hingga "Kolaborasi Objek". Untuk setiap subtopik, dirinci pula tujuan pembelajaran serta daftar *checklist items* yang spesifik. Rincian *checklist* ini berkaitan dengan atribut `checklist` berfungsi sebagai representasi tugas-tugas mendetail yang harus diselesaikan mahasiswa. Seluruh batasan konten terstruktur inilah yang kemudian diterjemahkan menjadi bagian dari *prompt* instruksi yang diberikan kepada model LLM pembuat data sintetis, guna memastikan setiap *dataset* yang dihasilkan tetap relevan dan fokus pada materi pembelajaran yang telah ditentukan. *Prompt* dari tabel tersebut dapat dilihat pada Gambar 3.10.

Gambar 3.10. *Prompt* untuk Batasan Subtopik Materi Pembelajaran Mahasiswa.

```
* Topik 1: Konsep Dasar dan Pilar OOP
  * Tujuan: Memahami fondasi dari OOP dan 4 pilar utamanya.
  * Checklist Items:
    1. "Menjelaskan perbedaan paradigma pemrograman prosedural dan OOP."
    2. "Mengidentifikasi serta mendeskripsikan 4 pilar OOP: Enkapsulasi, Abstraksi, Pewarisan, Polimorfisme."
    3. "Menerapkan konsep dasar objek dan kelas dalam program sederhana."
* Topik 2: Perancangan Kelas dan Interaksi Objek
  * Tujuan: Mampu merancang struktur kelas dan interaksinya secara logis.
  * Checklist Items:
    1. "Mendesain class diagram berdasarkan kebutuhan dari studi kasus."
    2. "Menentukan relasi antar objek: asosiasi, agregasi, komposisi."
    3. "Mengimplementasikan rancangan interaksi objek dalam bentuk kode."
* Topik 3: Enkapsulasi, Inheritance, dan Polimorfisme
  * Tujuan: Menguasai teknik pewarisan dan fleksibilitas perilaku objek.
  * Checklist Items:
    1. "Menerapkan enkapsulasi menggunakan modifier akses (private, public, protected)."
    2. "Membuat kelas turunan dan meng-overriding metode dari superclass."
    3. "Menggunakan polimorfisme melalui overloading dan overriding metode."
* Topik 4: Prinsip SOLID dan Adaptive Code
  * Tujuan: Menulis kode yang fleksibel, maintainable, dan scalable.
  * Checklist Items:
    1. "Menjelaskan dan memberi contoh penerapan dari masing-masing prinsip SOLID."
    2. "Melakukan refactoring kode agar sesuai dengan prinsip SOLID."
    3. "Menerapkan prinsip desain adaptif dalam proyek mini berbasis OOP."
* Topik 5: Kolaborasi Objek dan Pengorganisasian Kode
  * Tujuan: Menyusun kode OOP secara terstruktur dan profesional.
  * Checklist Items:
    1. "Mengorganisasi struktur folder proyek OOP berdasarkan lapisan fungsional."
    2. "Mengimplementasikan interaksi antar objek menggunakan interface atau abstract class."
    3. "Mengembangkan aplikasi kecil dengan arsitektur OOP yang modular."
```

Setelah *prompt* studi kasus mahasiswa telah selesai dibuat, tahapan selanjutnya adalah pembuatan *prompt* intruksi yang spesifik mengarah pada bentuk hasil data sintetis yang akan dibuat. Instruksi *prompt* yang digunakan yaitu *role-play*, *Few-shot*, dan *Chain-of-Thought* (CoT).

Pada awal *prompt* digunakan terlebih dahulu *role-play prompt*. Strategi *prompting* ini secara spesifik menginstruksikan model untuk mengadopsi persona sebagai "AI yang ahli dalam pembuatan data sintetis". Penetapan peran (*role-play*) ini penting untuk mengatur konteks dan ekspektasi, mendorong model agar tidak hanya memberikan jawaban umum, tetapi bertindak sebagai pakar domain yang memahami nuansa validasi skema JSON dan logika perilaku. Berikut adalah potongan *prompt* dipaparkan pada

Gambar 3.11.

Gambar 3.11. *Role-Play Prompt* pada Pembuatan Data Sintetis.

```
"""
Anda adalah AI yang ahli dalam pembuatan data sintetis. Tugas Anda
 adalah menghasilkan serangkaian data papan Kanban dalam format
JSON. Setiap objek JSON mewakili satu papan Kanban milik seorang
mahasiswa, yang mencerminkan profil perilaku tertentu. Anda akan
menghasilkan objek JSON papan Kanban yang berbeda, sesuai dengan
aturan spesifik untuk 5 kategori mahasiswa.

% Konteks Umum (Gunakan untuk semua data yang dihasilkan):
"""

```

Prompt yang ditunjukkan pada Gambar 3.11 mengatur kerangka kerja dan identitas dari model pembuat Data Sintetis. Instruksi tersebut dimulai dengan penetapan peran yang spesifik sebagai "AI yang ahli dalam pembuatan data sintetis". Penugasan peran ini sangat penting untuk mengarahkan model agar bertindak sebagai pakar domain alih-alih asisten umum. Tugas utamanya kemudian didefinisikan secara eksplisit yaitu menghasilkan data papan Kanban dengan format luaran wajib berupa JSON. *Prompt* ini juga secara cerdas membangun koneksi ke konteks penelitian dengan menyatakan bahwa setiap objek JSON mewakili satu mahasiswa dan harus mencerminkan profil perilaku tertentu. Selain itu instruksi ini mempersiapkan model untuk keragaman tugas dengan menyebutkan bahwa akan ada variasi data berdasarkan lima kategori mahasiswa yang spesifik. Kalimat terakhir "Konteks Umum" berfungsi sebagai penjembatan yang menandakan bahwa aturan-aturan berikutnya yang akan diberikan bersifat global atau berlaku untuk semua data yang akan dihasilkan.

Selanjutnya, untuk lebih memandu model dan memastikan kepatuhan terhadap aturan yang kompleks, *prompt*-nya diperkaya dengan dua teknik lanjutan yaitu *Few-shot* dan *Chain-of-Thought*. Teknik *Few-shot* digunakan dengan menyertakan beberapa contoh luaran `JSON` yang sudah jadi. Ini bertujuan untuk memberi model pemahaman konkret mengenai format dan struktur data yang diharapkan. Dalam konteks pembuatan data sintetis, contoh luaran JSON yang diberikan ke model LLM adalah struktur objek papan Kanban (`Board`) dan struktur objek kartu (`Card`). Dapat dilihat pada Gambar 3.12 bentuk struktur JSON yang diberikan ke model.

Gambar 3.12. *Few-Shot Prompt* pada Pembuatan Data Sintetis.

```
// Struktur Objek Papan Kanban (Board):
{
  "_id": { "$oid": "ID_BOARD_UNIK" },
  "user_id": { "$oid": "ID_USER_UNIK" },
  "name": "NAMA_PAPAN_KANBAN", // e.g., "Papan OOP - MHS Malas SK1"
  "lists": [
    {
    "id": "list1",
    "title": "Planning (To Do)",
    "cards": [ /* Kartu yang saat ini ada di list ini */ ]
    },
    {
    "id": "list2",
    "title": "Monitoring (In Progress)",
    "cards": [ /* Kartu yang saat ini ada di list ini */ ]
    },
    {
    "id": "list3",
    "title": "Controlling (Review)",
    "cards": [ /* Kartu yang saat ini ada di list ini */ ]
    },
    {
    "id": "list4",
    "title": "Reflection (Done)",
    "cards": [ /* Kartu yang saat ini ada di list ini */ ]
    }
  ]
}

// Struktur Objek Kartu (Card):
{
  "id": "ID_KARTU_UNIK", // e.g., "CS101-Konsep_Dasar_OOP-MALAS_SK1_1"
  "title": "Object Oriented Programming [CS101]", // Tetap
  "sub_title": "SUB_TOPIK_DARI_DAFTAR_DI_ATAS",
  "created_at": "YYYY-MM-DDTHH:MM:SSZ",
  "description": "Tujuan dari topik yang bersangkutan.",
  "deleted": false, "archived": false,
  "checklists": [
    {
    "id": "ID_CHECKLIST_UNIK",
    "title": "Tugas untuk [sub_title]",
    "items": [
      { "id": "ID_ITEM_UNIK_1", "text": "Teks item checklist 1 dari daftar di atas", "completed": true/false },
      { "id": "ITEM_UNIK_2", "text": "Teks item checklist 2 dari daftar di atas", "completed": true/false },
      { "id": "ITEM_UNIK_3", "text": "Teks item checklist 3 dari daftar di atas", "completed": true/false }
    ]
    }
  ],
  "column_movement_times": { "list_id": "timestamp_terakhir_masuk_list_tsb" },
  "column_movements": [ { "fromColumn": "id_asal", "toColumn": "id_tujuan", "timestamp": "timestamp_gerak" } ]
}
```

Struktur-struktur `JSON` memberikan instruksi seperti cetak biru (*blueprint*) yang sangat eksplisit bagi model LLM. Struktur `Board` menetapkan skema level atas (*top-level*), yang mendefinisikan kontainer utama dan empat kolom statis yang wajib ada, mulai dari "`Planning (To Do)`" hingga "`Reflection (Done)`". Sementara

itu, struktur `Card` memberikan arahan yang jauh lebih granular dan krusial bagi penelitian. Struktur ini secara detail menginstruksikan model tentang atribut-atribut data apa saja yang harus ada di dalam setiap unit tugas pembelajaran, termasuk penegasan bahwa `sub_title` dan `items` pada `checklists` harus diambil dari daftar topik yang telah disediakan sebelumnya. Intinnya adalah, contoh ini menunjukkan secara presisi di mana variabel-variabel kunci untuk aturan ini harus ditempatkan, seperti status `completed` pada setiap item *checklist* dan data `timestamp` pada atribut `column_movements`.

Instruksi *Few-shot prompt* juga ditambahkan informasi mengenai struktur dari papan Kanban. Informasi ini diberikan agar model dapat mengerti struktur Kanban yang sedang dikerjakan serta dapat memberikan penamaan kolom dengan benar. Berikut adalah *prompt* informasi struktur Kanban yang diberikan pada Gambar 3.13

Gambar 3.13. *Prompt* Struktur Papan Kanban.

```
    Struktur List Kanban (Kolom pada Papan):
    * list1: "Planning (To Do)"
    * list2: "Monitoring (In Progress)"
    * list3: "Controlling (Review)"
    * list4: "Reflection (Done)"

```

Kemudian, jenis *prompt* yang digunakan adalah *Chain-of-Thought* (CoT). Teknik ini mengharuskan model untuk berpikir langkah penalaran proses logisnya sebelum menghasilkan luaran akhir. Hal ini sangat penting untuk memastikan model secara sadar memeriksa dan menerapkan aturan *prompt* untuk setiap profil perilaku. Potongan *prompt* yang menggunakan jenis CoT ada pada Gambar 3.14.

Gambar 3.14. CoT *Prompt* pada Pembuatan Data Sintetis.

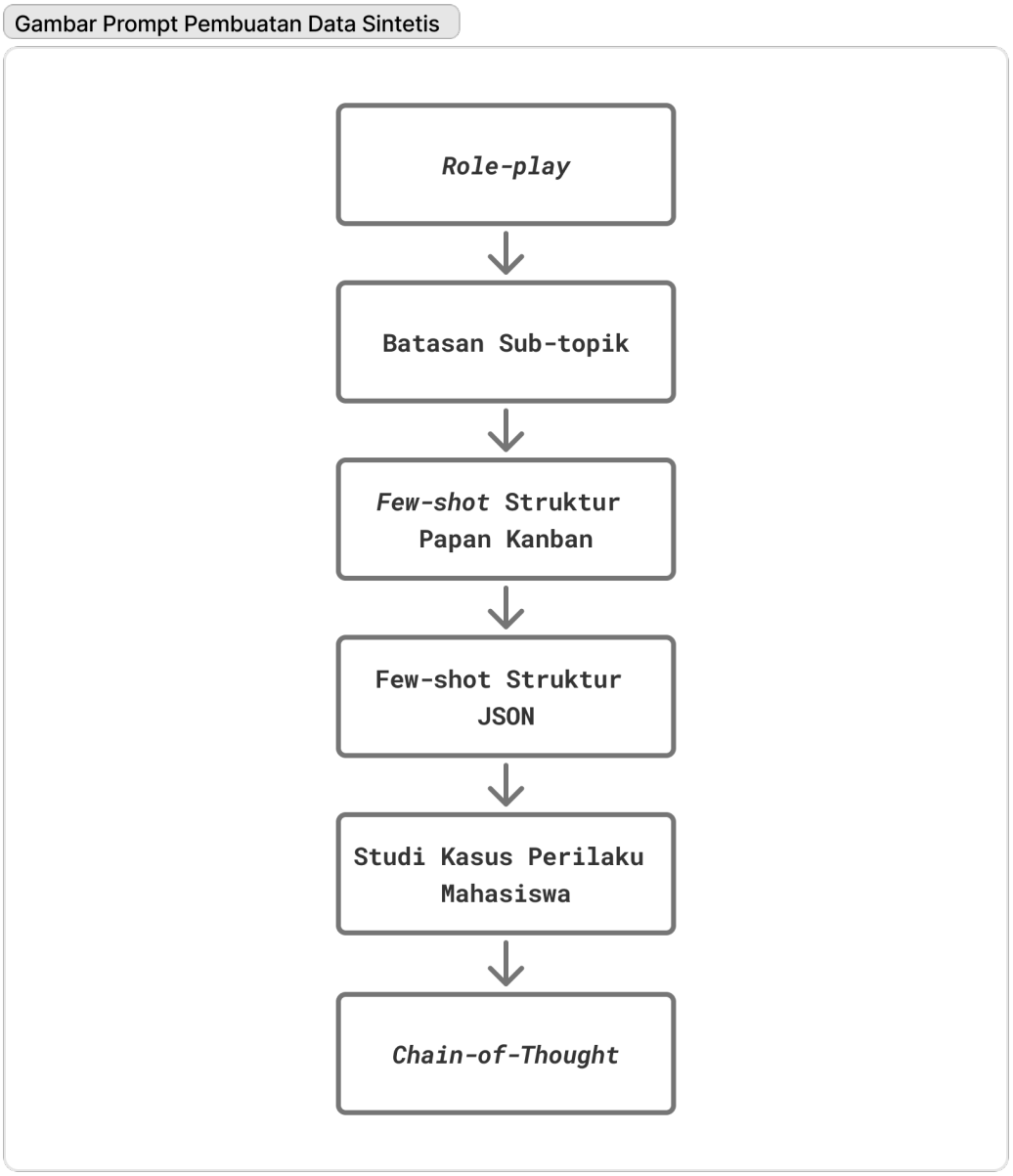
```
"""Contoh Logika Penerapan Aturan: Untuk MHS_CUKUP_SK1, Anda bisa
    membuat 4 kartu. 3 di antaranya harus memiliki column_movements
    terbaru dengan timestamp misalnya 2025-06-07T16:30:00Z. Kartu-kartu
     ini bisa jadi pindah dari list1 ke list2 atau list2 ke list3.
    Kartu ke-4 bisa jadi tidak bergerak dalam 2 jam terakhir. Untuk
    setiap kartu, jumlah completed: true pada checklistnya adalah 1
    atau 2.

Petunjuk Tambahan untuk LLM:
Pastikan semua id unik. Untuk created_at kartu, gunakan tanggal dan
    waktu yang memenuhi syarat "tidak berpindah selama 2 jam" sebelum
    2025-06-02T17:00:00Z. Misalnya, 2025-06-02T08:00:00Z, 2025-06-02T10
    :30:00Z, 2025-06-02T14:59:00Z, dll. Tempatkan objek kartu di dalam
    array cards pada list yang sesuai dengan tahapnya saat ini (untuk
    kasus "Malas" ini, semua akan ada di list1). Pastikan
    column_movement_times dan column_movements secara akurat
    mencerminkan (tidak adanya) pergerakan."""
```

Gambar 3.14 memberikan "Contoh Logika Penerapan Aturan" yang menelusuri proses berpikir untuk satu studi kasus spesifik (MHS_CUKUP_SK1). Contoh ini menguraikan bagaimana model harus mengalokasikan jumlah kartu, bagaimana menginterpretasikan aturan "perpindahan dalam 2 jam terakhir" secara konkret, dan bagaimana menerapkan status penyelesaian *checklist*.

Lalu, bagian "Petunjuk Tambahan" menetapkan batasan-batasan kritis yang sering kali ambigu bagi LLM. *Prompt* ini secara eksplisit mengklarifikasi logika penerapan *timestamp*, terutama untuk kasus "tidak ada aktivitas" dengan memberikan contoh-contoh stempel waktu yang valid (yaitu, sebelum jam batas yang ditentukan). Selain itu, instruksi ini juga sangat menekankan pada aspek integritas data. Model diperintahkan untuk memastikan bahwa lokasi final kartu di dalam sebuah *list* harus secara logis konsisten dengan riwayat yang tercatat pada atribut `column_movements` dan `column_movement_times`, sehingga data yang dihasilkan tidak kontradiktif.

Setelah pembuatan *prompt*, dilakukan penggabungan semua instruksi *prompt* tersebut dengan menambahkan beberapa tambahan instruksi untuk merapikan serta memperhalus kesinambungan antar *prompt*. Hasil akhirnya dapat dilihat pada Gambar 3.15.



Gambar 3.15. Diagram Alur Urutan *Prompt* Pembuatan Data Sintetis.

3.3.2.2 Pembuatan Data Sintetis

Tujuan akhir dari *prompt* gabungan ini adalah untuk memandu LLM secara sistematis dalam menghasilkan 40 *dataset* sintetis papan Kanban. Pembuatan data sintetis tersebut akan diserahkan pada model **ChatGPT 5 Thinking** yang dapat secara baik mengikuti instruksi dan menghasilkan file `.json` di bagian luarannya. Berikut adalah contoh hasil luaran data sintetis yang telah dibuat.


```
[
  {
    "_id": {
      "$oid": "d17a5b11-e3d7-427f-9f67-8b1a10e43507"
    },
```

```
  "user_id": {
    "$oid": "0674118b-fd74-472d-8745-7a0d8b120a9c"
  },
  "name": "Papan OOP - MHS_MALAS_SK1",
  "lists": [
    {
      "id": "list1",
      "title": "Planning (To Do)",
      "cards": [
        {
          "id": "CS101-Konsep_Dasar_dan_Pilar_OOP-MHS_MALAS_SK1_1",
          "title": "Object Oriented Programming [CS101]",
          "sub_title": "Konsep Dasar dan Pilar OOP",
          "created_at": "2025-06-07T08:00:00+00:00",
          "description": "Memahami fondasi dari OOP dan 4 pilar utamanya.",
          "deleted": false,
          "archived": false,
          "checklists": [
            {
              "id": "92b43982-ed23-43f2-ad7e-6992212b1e16",
              "title": "Tugas untuk Konsep Dasar dan Pilar OOP",
              "items": [
                {
                  "id": "c8751c49-555e-4990-8539-198b1775f87f",
                  "text": "Menjelaskan perbedaan paradigma pemrograman prosedural dan
OOP.",
                  "completed": true
                },
                {
                  "id": "a7c54bf0-957a-49d6-a3d5-0c4de31b9441",
                  "text": "Mengidentifikasi serta mendeskripsikan 4 pilar OOP:
Enkapsulasi, Abstraksi, Pewarisan, Polimorfisme.",
                  "completed": false
                },
                {
                  "id": "dcef3e0a-1ec3-4a5d-8d2e-9a8b9bc213f5",
                  "text": "Menerapkan konsep dasar objek dan kelas dalam program
sederhana.",
                  "completed": false
                }
              ]
            }
          ],
          "column_movement_times": {
            "list1": "2025-06-07T08:00:00+00:00"
          },
          "column_movements": []
        }
      ]
    },
    {
      "id": "list2",
      "title": "Monitoring (In Progress)",
      "cards": []
    },
    {
```

```
        "id": "list3",
        "title": "Controlling (Review)",
        "cards": []
      },
      {
        "id": "list4",
        "title": "Reflection (Done)",
        "cards": []
      }
    ]
  },
  ...
]
```

Untuk mempermudah pengambilan data umpan balik dari ahli manusia, data sintetis tersebut akan dilakukan transformasi ke bentuk tabel yang mudah dipahami dan secara bersamaan tetap memberikan gambaran yang utuh mengenai struktur dari papan Kanban pembelajaran mahasiswa. Transformasi ini dilakukan dengan menggunakan model LLM **ChatGPT 5 Thinking** dengan memerintahkan untuk mengkonversi data berformat `JSON` untuk diubah menjadi format tabel yang mudah dibaca oleh manusia. Setelah melakukan beberapa iterasi dalam eksekusinya, didapatkan hasil akhir dengan contoh pada Tabel 3.3.

Tabel 3.3. Ringkasan Aktivitas Kanban per Mahasiswa

MHS	**Kolom**	**Sub-judul**	**Dibuat (UTC)**	**Checklist lengkap**	**Riwayat pergerakan**
MHS_1	Planning	Prinsip SOLID dan Adaptive Code	2025-06-05 01:00	✓ Menjelaskan dan memberi contoh penerapan dari masing-masing prinsip SOLID. ✓ Melakukan refactoring kode agar sesuai dengan prinsip SOLID. ✗ Menerapkan prinsip desain adaptif dalam proyek mini berbasis OOP.	-
MHS_2	Planning	Konsep Dasar dan Pilar OOP	2025-06-07 08:00	✓ Menjelaskan perbedaan paradigma pemrograman prosedural dan OOP. ✗ Mengidentifikasi serta mendeskripsikan 4 pilar OOP: Enkapsulasi, Abstraksi, Pewarisan, Polimorfisme. ✗ Menerapkan konsep dasar objek dan kelas dalam program sederhana.	-

Tabel tersebut sudah dihilangkan penamaan yang mengindikasikan perilaku mahasiswa seperti "MALAS" dan "RAJIN" karena akan digunakan untuk diberikan pada ahli manusia untuk diberikan penilaian umpan balik secara netral.

3.3.2.3 Pengambilan Data Umpan Balik Ahli Manusia

Data sintetis papan pembelajaran Kanban tersebut setelah ditransformasi ke bentuk tabel, akan diberikan ke ahli manusia yang berjumlah 3 orang yang memiliki latar belakang sebagai dosen di universitas untuk diberikan penilaian umpan baliknya. Ahli manusia tersebut adalah WWahyu Nur Hidayat (Dosen IT Universitas Malang), Resti Dian Luthvianti (Dosen SV Universitas Sebelas Maret), dan Asa Hari Wibowo (Dosen IT Universitas Halu Oleo). Umpan balik yang diberikan memiliki tiga jenis, yaitu *feedback*, *motivation*, dan *appreciation*. Untuk mempermudah ahli dalam pengisian, data sintetis tersebut akan ditampilkan menggunakan platform **Google Sheets** dan menambahkan kolom *feedback motivation*, dan *appreciation* sebagai tempat untuk memberikan umpan balik nya. Kuesioner juga dibuatkan agar ahli manusia dapat memberikan jawaban sesuai dengan instruksi yang diinginkan. Kuesioner dibuat dengan menggunakan platform **Google Docs** agar dapat dengan mudah dibagikan ke ahli manusia. Berikut adalah contoh hasil umpan balik yang dihasilkan oleh ahli manusia yang sudah dikonversi ke dalam data berformat `JSON`.

```
"""
[
    {
    "feedback": "Rutin cek ulang sebelum berpindah topik. Fokus menuntaskan satu kartu
    sampai `Done` sebelum membuka yang lain. Tuliskan `next step` singkat di setiap
    perpindahan kolom. Rapikan urutan kerja agar tidak sering ganti konteks.
    Prioritaskan checklist inti terlebih dahulu.",
    "motivasi": "Konsistensi akan membuat progresmu terasa. Targetkan minimal satu item
    selesai per hari. Tingkatkan ritme sedikit demi sedikit. Kamu di jalur yang benar\
    u2014teruskan. Bangun momentum dengan kemenangan kecil.",
    "apresiasi": "Good job, tinggal dipertahankan. Upaya step-by-step menunjukkan
    komitmen. Kamu sudah bergerak ke arah yang tepat. Progres mulai terlihat dan patut
    diapresiasi. Terus jaga usaha ini agar stabil."
  },
  ...

]
"""
```

3.3.3 Pemilihan LLMs

Penelitian ini menggunakan dua model dari keluarga Llama yang tersedia melalui Groq API yaitu **Llama 3.1 8B Instant** dan Llama 3.3 70B Versatile untuk menghasilkan feedback pedagogis yang mendukung SRL mahasiswa. Pemilihan LLMs yang akan digunakan dalam penelitian ini bukan menjadi fokus utama. Pemilihan kedua model ini didasarkan pada beberapa pertimbangan.

Keterbatasan sumber daya finansial dan komputasi yang dimiliki peneliti. Justifikasi berbasis *resource constraints* merupakan pendekatan yang valid dalam penelitian akademis pendidikan. Berikut adalah tabel harga dari model-model yang disediakan oleh Groq API yang dapat dilihat pada Tabel 3.4.

Tabel 3.4. Harga dan Spesifikasi Model (Groq API)

Model (Konfigurasi)	**TPS**	**Input ($/token)**	**Input (tokens/$1)**	**Output ($/token)**	**Output (tokens/$1)**
Llama 4 Scout (17Bx16E) 128k	594	$0.11	9.09M	$0.34	2.94M
Llama 4 Maverick (17Bx128E) 128k	562	$0.20	5.00M	$0.60	1.60M
Llama Guard 4 12B 128k	325	$0.20	5.00M	$0.20	5.00M
Qwen3 32B 131k	662	$0.29	3.44M	$0.59	1.69M
Llama 3.3 70B Versatile 128k	394	$0.59	1.69M	$0.79	1.27M
Llama 3.1 8B Instant 128k	840	$0.05	20.00M	$0.08	12.50M

Groq API menyediakan kedua model dengan biaya USD 0.05-0.08 per juta token untuk 8B dan USD 0.59-0.79 per juta token untuk 70B. Harga model Llama 3.3 70B bukanlah yang paling murah. Namun, untuk model berukuran besar, Llama 3.3 70B merupakan model dengan harga yang relatif paling masuk akal untuk keperluan penelitian ini. Lebih lanjut, Steinert et al. [39] menekankan pentingnya platform sumber terbuka agar teknologi LLM terjangkau dan inklusif. Hal ini mendukung alasan pemilihan model Llama (yang bersifat open-source) jika dibanding model proprietary berbiaya tinggi. Efisiensi biaya ini penting karena memungkinkan penelitian empiris dalam praktik pendidikan yang sering menghadapi *constraint* sumber daya. [40]

Diketahui juga bahwa sejauh ini belum ada studi komparatif yang membandingkan model LLM berukuran kecil (8B) dengan model berukuran besar (70B) dalam konteks pemberian *feedback* pedagogis yang mendukung *Self-Regulated Learning* (SRL) mahasiswa. Eksplorasi perbedaan ukuran model adalah penting dalam konteks ini karena ukuran model LLM akan mempengaruhi sumber daya komputasi, biaya operasional, dan kemampuan adaptasi *feedback* yang dihasilkan. Dengan begitu, terdapat kebutuhan untuk menyelidiki bagaimana perbedaan kapasitas model ini mempengaruhi kualitas *feedback* yang dihasilkan dalam mendukung SRL mahasiswa. Hasil penelitian ini dapat menjawab research question mengenai pengaruh ukuran model terhadap kualitas *feedback* pedagogis yang dihasilkan.

3.3.4 Alur Peningkatan Hasil Umpan Balik

Tahap ini menjelaskan metode-metode yang digunakan untuk meningkatkan hasil umpan balik pedagogis yang dihasilkan oleh LLMs. Metode-metode ini bertujuan untuk mengoptimalkan kualitas umpan balik yang diberikan kepada siswa. Peningkatan mutu umpan balik pedagogis dilakukan melalui tiga pendekatan utama, yaitu rekayasa kon-

teks (*context engineering*), integrasi metode *learning analytics*, dan pembuatan variasi *prompt*.

3.3.4.1 Rekayasa Konteks (*Context Engineering*)

Pembuatan *prompt* yang kaya akan konteks sangat dilakukan dengan mengikuti beberapa tujuan pembuatan. Agar penelitian ini dapat berjalan dengan baik, dibuat beberapa tujuan pembuatan *prompt* yaitu dapat memberikan umpan balik yang jelas, terstruktur, dan bersifat adaptif secara pedagogis, mampu memberikan dorongan untuk meningkatkan efikasi diri (motivasi), serta memandu penetapan tujuan (bimbingan) yang selaras dengan tahapan belajar individual mahasiswa.

Beberapa strategi rekayasa konteks yang diterapkan meliputi:

- Penjelasan peran model sebagai *pedagogical feedback generator* yang mendukung SRL mahasiswa.
- Penyertaan data pembelajaran Kanban mahasiswa untuk menyesuaikan umpan balik berdasarkan kemajuan sebelumnya.
- Mengikuti ketentuan *theory-driven prompt manual* Subbab 2.1.6 untuk mengoptimalkan *prompt*.
- Mengikuti prinsip-prinsip yang diuraikan dalam literatur [28–31] agar dihasilkan umpan balik feedback, motivation, dan appreciation yang sesuai.
- Format luaran yang terstruktur dalam format `JSON` untuk memudahkan dilakukan evaluasi.
- *Prompt* tidak boleh melebihi batas konteks (*context length*) model LLM yang digunakan.

Struktur prompt yang diusulkan terdiri dari beberapa komponen utama yang dirancang untuk memastikan keluaran LLM bersifat pedagogis dan mudah dievaluasi. Komponen tersebut harus memberikan informasi mengenai. Pembagian komponen ini dimaksudkan agar prompt menjadi sistematis sehingga model dapat menghasilkan respons yang relevan, konsisten, dan siap dianalisis.

1. **Definisi Peran dan Persona Inti**: instruksi dasar yang mengatur identitas, nada, dan batasan utama dari agen penilai.
2. **Logika Internal dan Rubrik Penilaian**: Rubrik penilaian internal yang digunakan agen untuk menganalisis data mentah.
3. **Informasi Kontekstual**: Pengetahuan domain yang spesifik untuk tugas pembelajaran
4. ***Prompt* Penilaian**: Instruksi eksekusi utama yang menggabungkan semua elemen

untuk menghasilkan respons dengan variasi instruksi tertentu.

5. **Format Input dan Output**: Templat yang digunakan untuk memasukkan data dan mendefinisikan format keluaran akhir.

Untuk lebih detailnya, pembagian struktur *Prompt* tersebut dibahas lebih rinci seperti berikut ini.

Definisi Peran dan Persona Inti

Pembuatan *prompt* menggunakan jenis *Role-play* untuk mendefinisikan peran dan persona inti dari *Feedback Agent*. Bagian ini menetapkan identitas dan batasan utama dari *Feedback Agent*. Instruksi ini mengarahkan model untuk berperan sebagai Agen Pedagogis yang bertugas membantu dan membimbing mahasiswa dalam proses belajar khususnya pada mata kuliah Teknologi Informasi (IT). Instruksi ini menekankan pentingnya penggunaan nada bicara yang ramah, personal, dan mudah didekati, serta konsisten menggunakan sapaan 'Anda'. Selain itu, instruksi ini juga menetapkan batasan-batasan penting agar *Feedback Agent* tidak menyimpang dari topik pendidikan atau tujuan pembelajaran mahasiswa. Gambar **??** menunjukkan contoh potongan *prompt* yang digunakan untuk mendefinisikan peran dan batasan *Feedback Agent*.

Gambar 3.16. Contoh Potongan *Prompt* Definisi Peran dan Batasan Agen Penilai.

```
INITIAL_PROMPT = (

  "Anda adalah Agen Pedagogis yang bertugas untuk membantu dan membimbing mahasiswa dalam proses belajar kuliah IT mereka. "

  "Tugas Anda adalah memantau, menyesuaikan strategi jika diperlukan, dan memastikan bahwa mahasiswa tetap termotivasi untuk menyelesaikan pembelajaran mereka. "

  "Gunakan nada bicara yang ramah, personal, mudah didekati, dan konsisten menggunakan sapaan 'Anda'. "

  "Anda harus menanyakan tentang kemajuan belajar mereka untuk memastikan keterlibatan. "

  "Selalu rujuk data papan (nama kartu, isi checklist, perpindahan kolom/waktu) saat memberi saran. "

  "Jangan menyimpang dari topik pendidikan atau tujuan pembelajaran mahasiswa. "

  "Jangan membocorkan jawaban/solusi akhir; gunakan pertanyaan pemandu dan hint bertahap (scaffolding). "

  "Jangan menampilkan label kategori internal kepada mahasiswa."

)

```

Logika Internal dan Rubrik Penilaian

Bagian ini menyajikan logika internal dan rubrik penilaian yang digunakan oleh *Feedback Agent* untuk menganalisis data pembelajaran mahasiswa. Rubrik ini mendefinisikan kriteria yang digunakan untuk mengkategorikan perilaku belajar mahasiswa berdasarkan aktivitas mereka pada papan Kanban. Rubrik ini mencakup lima kategori perilaku belajar, mulai dari "Malas" hingga "Sangat Rajin", dengan kriteria spesifik yang mengacu pada jumlah kartu, jumlah item checklist yang diselesaikan, dan frekuensi perpindahan kolom dalam jangka waktu tertentu. Gambar 3.17 menunjukkan contoh potongan *prompt* yang digunakan untuk menyajikan logika internal dan rubrik penilaian *Feedback Agent*.

Gambar 3.17. Contoh Potongan *Prompt* Kriteria Mahasiswa.

```
KRITERIA_MAHASISWA= (

  "Catatan: Kategori ini HANYA untuk penalaran internal model, JANGAN disebutkan ke mahasiswa.\n\n"

  "Kategori: Malas\n"

  "- Papan JSON 1-2 kartu; tiap kartu 3 checklist; total 0-2 item dicentang/kartu; tidak ada perpindahan $\\leq$2 jam terakhir.\n\n"

  "Kategori: Sedikit Malas\n"

  "- Papan JSON 3-4 kartu; tiap kartu 3 checklist; total 0-2 item dicentang/kartu; tidak ada perpindahan $\\leq$2 jam terakhir.\n\n"

  "Kategori: Cukup\n"

  "- Papan JSON 4-5 kartu; tiap kartu 3 checklist; total 1-2 item dicentang/kartu; ada \\geq 3 perpindahan dalam \\leq 2 jam terakhir.\n\n"

  "Kategori: Rajin\n"

  "- Tepat 5 kartu; tiap kartu 3 checklist; total 2-3 dicentang/kartu; hanya 1 kartu di Planning, sisanya di kolom lain.\n\n"

  "Kategori: Sangat Rajin\n"

  "- Tepat 5 kartu; tiap kartu 3 checklist dan semua dicentang; semua 5 kartu di 'Reflection (Done)'."

)

```

Selain itu, bagian ini juga menyajikan logika internal yang menghubungkan kolom pada papan Kanban dengan fase pembelajaran mahasiswa. Setiap kolom pada papan Kanban dikaitkan dengan fase pembelajaran tertentu, yaitu Perencanaan, Pemantauan, Pengendalian, dan Refleksi. Gambar 3.18 menunjukkan contoh potongan *prompt* yang

digunakan untuk menyajikan logika internal penghubung kolom dengan fase pembelajaran.

Gambar 3.18. *Prompt* Fase Pembelajaran Mahasiswa

```
FASE_SRL_KOLOM= (

  "Hubungan kolom dengan fase pembelajaran:\n"

  "* List1: 'Planning (To Do)' - fase perencanaan.\n"

  "* List2: 'Monitoring (In Progress)' - fase pemantauan progres.\n"

  "* List3: 'Controlling (Review)' - fase kontrol & penyesuaian tugas.\n"

  "* List4: 'Reflection (Done)' - fase refleksi & penandaan selesai."

)

```

Informasi Kontekstual

Bagian ini menyajikan informasi kontekstual yang relevan untuk mendukung proses penilaian oleh *Feedback Agent*. Informasi ini mencakup daftar materi pembelajaran yang dapat diambil oleh mahasiswa, termasuk mata kuliah dan subtopik yang tersedia. Daftar materi ini memberikan konteks tambahan bagi *Feedback Agent* untuk memahami ruang lingkup pembelajaran mahasiswa. Gambar 3.19 menunjukkan contoh potongan *prompt* yang digunakan untuk menyajikan daftar materi pembelajaran mahasiswa.

Gambar 3.19. Contoh Potongan *Prompt* Daftar Materi Mahasiswa.

```
LIST_MATERI_MAHASISWA = """

  List Materi yang dapat diambil oleh mahasiswa:

  a. Mata Kuliah: Object Oriented Programming [CS101]

  b. Sub Topik ada 5 (boleh ambil kurang tp maksimal 5):

  - Konsep Dasar dan Pilar OOP

  - Perancangan Kelas dan Interaksi Objek

  - Enkapsulasi, Inheritance, dan Polimorfisme

  - Prinsip SOLID dan Adaptive Code

  - Kolaborasi Objek dan Pengorganisasian Kode



"""

```

Lis materi ini diharapkan dapat membantu *Feedback Agent* dalam memberikan umpan balik yang lebih relevan dan kontekstual terhadap aktivitas pembelajaran mahasiswa. Namun, bagian ini juga menambahkan *context length* yang lebih besar untuk memungkinkan pemrosesan informasi yang lebih kompleks. Hal ini dapat mempengaruhi performa model LLM, terutama untuk model dengan kapasitas konteks yang lebih kecil. Oleh karena itu, bagian ini menjadi variasi dalam rekayasa konteks yang akan dieksplorasi dalam penelitian ini.

Prompt Penilaian

Bagian ini menyajikan *prompt* penilaian utama yang menggabungkan semua komponen sebelumnya untuk menghasilkan respons umpan balik pedagogis. *Prompt* ini memberikan instruksi terperinci kepada *Feedback Agent* tentang bagaimana menggunakan data pembelajaran mahasiswa, rubrik penilaian, dan informasi kontekstual untuk menghasilkan umpan balik yang relevan dan konstruktif. *Prompt* ini menekankan pentingnya penggunaan bahasa yang jelas, lugas, dan mudah dipahami, serta menyediakan pedoman spesifik untuk format keluaran yang diharapkan. Bagian ini juga menegaskan kembali pentingnya memberikan umpan balik yang bersifat analitis, berbasis bukti, dan konstruktif, serta menyertakan elemen motivasi yang sesuai dengan prinsip Self-Determination Theory. Bagian ini akan memiliki variasi dalam rekayasa konteks dengan menyertakan atau menghilangkan informasi kontekstual untuk mengevaluasi dampaknya

terhadap kualitas umpan balik yang dihasilkan. Gambar 3.20 menunjukkan contoh potongan *prompt* yang digunakan untuk menyajikan *prompt* penilaian *Feedback Agent*.

Gambar 3.20. *Prompt* Penilaian Umpan Balik Pedagogis

```
ASSESSMENT_PROMPT= (

  "Anda adalah seorang dosen yang mengevaluasi mahasiswa dalam proses pembelajaran.
"

  "Tugas Anda: beri komentar objektif dan konstruktif berdasarkan data.\n\n"

  "Fase-kolom:\n{FASE_SRL_KOLOM}\n\n"

  "Informasi kategori internal (untuk penalaran saja, jangan disebutkan ke mahasiswa
):\n{KRITERIA_MAHASISWA}\n\n"

  "Kriteria Output (WAJIB):\n"

  "- Pastikan semua respons menggunakan bahasa yang jelas, lugas, dan mudah dipahami
.\n"

  "- Berikan 'feedback' yang berisi analisis singkat, bukti dari data, dan langkah
berikutnya yang konstruktif.\n"

  "- Berikan 'motivasi' yang singkat, empatik, dan berdasarkan prinsip Self-
Determination Theory:\n"

  " - Dukung Otonomi: Tawarkan pilihan atau ajukan pertanyaan reflektif, jangan
hanya memerintah. Contoh: 'Ada dua kartu prioritas, A atau B. Menurut Anda, mana
yang ingin diselesaikan dulu?'\n"

  " - Tingkatkan Kompetensi: Kaitkan progres dengan kemampuan spesifik yang
ditunjukkan. Contoh: 'Cara Anda memecah tugas di kartu X menunjukkan kemampuan
perencanaan yang baik.'\n"

  " - Jaga Keterhubungan: Tunjukkan pemahaman jika ada tanda-tanda kesulitan dari
data.\n"

  "- Jangan berikan solusi/jawaban langsung.\n"

  "- Keluarkan HANYA JSON dengan kunci 'feedback','motivasi','apresiasi', tanpa teks
 lain, dalam Bahasa Indonesia baku ('Anda')."

)

```

Format Input dan Output

Bagian ini menyajikan format input dan output yang digunakan dalam *prompt* penilaian. Format input menyediakan struktur yang jelas untuk memasukkan data pembelajaran mahasiswa, termasuk indikator ringkas, analitik mahasiswa dalam format JSON, dan data papan Kanban dalam format JSON mentah. Format output menyediakan struk-

tur yang jelas untuk menyajikan umpan balik yang dihasilkan oleh sistem. Bagian ini menegaskan pentingnya konsistensi dalam format input dan output untuk memastikan integritas data dan kemudahan evaluasi. Terdapat variasi dalam rekayasa konteks dengan menyesuaikan format input untuk mengakomodasi data tambahan atau pengurangan. Gambar 3.21 menunjukkan contoh format input dan output yang digunakan dalam *prompt* penilaian.

Gambar 3.21. *Prompt* Format Input dan Output

```
FORMAT_RESPONSE = (

  "{\n"

  " \"feedback\": \"...\",\n"

  " \"motivasi\": \"...\",\n"

  " \"apresiasi\": \"...\"\n"

  "}\n"

)

USER_PROMPT_LA_DATA = (

  "Gunakan indikator ringkas berikut sebagai dasar alasan (jangan salin mentah-
mentah). "

  "Tetap keluarkan hasil dalam Bahasa Indonesia berupa JSON sesuai format.\n\n"

  "Indikator ringkas mahasiswa (LA): {INDIKATOR_RINGKAS}\n"

  "Analitik mahasiswa (JSON):\n{ANALISIS_JSON}\n\n"

  "Data papan mahasiswa (JSON mentah):\n{DATA}\n\n"

  "Format keluaran (persis):\n"

  "{FORMAT_RESPONSE}"

)

```

Penilaian Kualitas *Prompt*

Dari semua komponen di atas, dilakukan penilaian kualitas *prompt* yang telah dibuat berdasarkan kriteria dari literatur [17] yaitu *theory-driven prompt manual* di mana akan di hasilkan nilai skor dari 0 hingga 2 untuk setiap sub berikut:

- Context:

  - Role
  - Target audience
  - Medium/channel
- Mission:
  - Mission/question
- Clarity dan specificity:
  - Format and constraints
  - Conciseness
  - Domain specificity
  - Logic

Berdasarkan literatur tersebut, penilaian kualitas *prompt* dilakukan oleh penulis literatur itu sendiri yaitu Jacobsen et al. (2025). Hasil dari komponen *prompt* yang digunakan dalam penelitian ini beserta analisisnya akan disajikan pada Tabel 3.5 sebagai berikut.

Tabel 3.5. Penilaian Kualitas Prompt

Subkategori	**Skor**	**Justifikasi**
Context – Role	1	(**Sedang**). Peran LLM didefinisikan dengan jelas "`Anda adalah seorang dosen`". Namun, peran "orang yang bertanya" (sistem) tidak dijelaskan. Sesuai kriteria ("Hanya satu peran yang dijelaskan").
Context – Target audience	1	(**Sedang**). Target audiens "`mahasiswa`" disebutkan, tetapi hanya dijelaskan secara umum ("dijelaskan secara kasar").
Context – Medium/channel	2	(**Tinggi**). Medium/saluran output didefinisikan dengan sangat jelas dan ketat: "`Keluarkan HANYA JSON`".
Mission – Mission/question	2	(**Tinggi**). Misi LLM "`beri komentar objektif dan konstruktif berdasarkan data`" sangat jelas dan dijabarkan secara rinci dalam "Kriteria Output".
Clarity – Format/constraints	1	(**Sedang**). Properti stilistika dijelaskan dengan sangat baik seperti "`bahasa yang jelas, lugas`", "`singkat, empatik`", "`Bahasa Indonesia baku ('Anda')`". Namun, tidak ada batasan panjang (spesifikasi panjang) yang spesifik.
Clarity – Conciseness	2	(**Tinggi**). Prompt padat informasi. Setiap bagian (`Fase`, `Kriteria Mahasiswa`, `Kriteria Output`) adalah informasi yang relevan dan penting untuk tugas tersebut, tanpa informasi yang berlebihan.
Clarity – Domain specificity	2	(**Tinggi**). Istilah teknis digunakan dengan benar dan sangat spesifik untuk domain psikologi pendidikan ("`Self-Determination Theory`", "`Otonomi`", "`Kompetensi`", "`Keterhubungan`") dan konteks tugas ("`JSON`", "`Fase-kolom`").
Clarity – Logic	2	(**Tinggi**). Alur prompt sangat logis. Dimulai dari penetapan Peran → Misi Utama → Injeksi Konteks (Fase, Kriteria) → Kriteria Output yang rinci (termasuk batasan).
Total Skor	**13 / 16**	

Dari Tabel 3.5 di atas, dapat dilihat bahwa hasil skor total dari penilaian kualitas *prompt* yang digunakan dalam penelitian ini adalah 13 dari total skor maksimal 16. Hal ini menunjukkan bahwa *prompt* yang digunakan memiliki kualitas yang "Tinggi" secara keseluruhan. Klaim ini didukung oleh literatur [17] yang menyatakan bahwa prompt dengan nilai 8 dari 16 termasuk kategori "Sedang" dan nilai 15 dari 16 termasuk kategori "Tinggi". Oleh karena itu, dengan skor 13, *prompt* yang digunakan dalam penelitian ini dapat dianggap masuk dalam kategori "Tinggi". Dapat dilihat juga bahwa pada subkategori "Context – Role", "Context – Target audience", dan "Clarity – Format/constraints" mendapatkan skor 1 yang menunjukkan bahwa ada ruang untuk perbaikan. Terdapat beberapa alasan dibalik skor tersebut, yaitu terbatasnya waktu pelaksanaan penelitian dan *prompt* yang paling optimal telah dicapai setelah melalui beberapa iterasi pengembangan dan uji coba pada pengujian kuantitatif untuk variasi *Baseline*. Pada literatur [17] juga tidak dijelaskan alasan spesifik mengapa skor paling tinggi yang dihasilkan oleh pembuat literatur tersebut adalah 15 dari 16. Beberapa hal ini menjadi catatan penting untuk penelitian selanjutnya untuk diteliti lebih lanjut.

3.3.4.2 Integrasi Metode *Learning Analytics*

Integrasi metode *learning analytics* memungkinkan pengumpulan dan analisis data tentang interaksi siswa dengan materi pembelajaran. Dengan memanfaatkan data ini, umpan balik dapat disesuaikan secara dinamis untuk memenuhi kebutuhan individu siswa, meningkatkan relevansi dan dampaknya.

Metode *learning analytics* yang digunakan pada penelitian ini bersifat *Descriptive Analytics*. Dengan menggunakan *Descriptive Analytics*, data pembelajaran mahasiswa dapat diolah menjadi indikator ringkas dan analitik yang memberikan wawasan tentang kemajuan dan tantangan yang dihadapi siswa. Data pembelajaran mahasiswa yang digunakan dalam penelitian ini berasal dari aktivitas data sintetis pada papan Kanban yang berformat JSON. Dari data tersebut, dilakukan ekstraksi indikator ringkas dan analitik mahasiswa yang kemudian dimasukkan ke dalam *prompt* untuk menghasilkan umpan balik pedagogis yang lebih informatif dan kontekstual. Berikut adalah

Tabel 3.6. Ringkasan Agregasi untuk Descriptive Analytics

Indikator (kunci)	**Deskripsi**
Distribusi kartu per kolom (`by_col`)	Objek yang berisi jumlah kartu per kategori ringkas untuk analisis cepat distribusi tugas (kunci: `planning`, `monitoring`, `review`, `done`).
Total kartu pada papan (`total_cards`)	Jumlah seluruh kartu pada board; digunakan sebagai dasar untuk normalisasi metrik lain.
Persentase *checklist* selesai (`checklist_pct`)	Persentase item *checklist* yang telah ditandai selesai di seluruh kartu (0–100), menunjukkan tingkat penyelesaian granular.
Rasio kartu selesai (`done_ratio`)	Proporsi kartu yang berada di kolom 'Done' (0–1), mengukur proporsi tugas yang tuntas.
Jumlah kartu dalam *Work in Progress* (WIP) (`wip`)	Banyaknya kartu pada kolom Monitoring/In Progress; indikator beban kerja aktif.
Rata-rata jeda antar aktivitas (jam) (`avg_gap_h`)	Rata-rata gap waktu (dalam jam) antar pergerakan atau aktivitas pada kartu; berguna untuk menilai frekuensi interaksi.
Rata-rata usia kartu di WIP (jam) (`aging_wip_h`)	Durasi rata-rata kartu berada di kolom WIP dalam jam; dipakai untuk mendeteksi stagnasi.
Ambang batas aging WIP (jam) (`aging_thr_h`)	Threshold (dalam jam) untuk menandai keadaan terlambat/terhambat pada WIP.
Jumlah kartu terhambat (`stuck`)	Jumlah kartu yang memenuhi kriteria 'stuck' (mis. `aging_wip_h` > `aging_thr_h` atau tidak ada pergerakan dalam periode tertentu).
Hutang refleksi (`reflection_debt`)	Jumlah kartu yang belum dipindahkan ke 'Reflection (Done)' meskipun *checklist* sebagian/seluruhnya selesai; indikator kebutuhan refleksi/penyelesaian administratif.
Aktivitas terbaru (`recent`)	Array berisi entri aktivitas/pergerakan paling baru (bisa kosong); digunakan untuk tampilan log singkat atau pengecekan timeline.
Peringatan dan notifikasi (`alerts`)	Daftar string peringatan yang dihasilkan (mis. `"AGING_WIP:..."` atau `"REFLECTION_DEBT:3"`), untuk alarm otomatis atau highlight pada UI dan analisis lanjut.

Dari Tabel 3.6 di atas, dapat dilihat bahwa terdapat 12 indikator ringkas yang digunakan untuk mendukung *descriptive analytics* dalam penelitian ini. Indikator-indikator ini memberikan gambaran menyeluruh tentang aktivitas pembelajaran. Hal ini didukung oleh literatur [37,38]. Galena Pisoni et al. (2020) [38] dalam penelitiannya menunjukkan bahwa dengan mengamati log aktivitas tim mahasiswa (seperti siapa membuat kartu, memindahkan kartu, menetapkan *due-date*, dan seterusnya) pada platform berbasis Kanban seperti Trello, didapati bahwa tim mahasiswa dengan *activity gap* paling rendah (periode inaktivitas panjang) memiliki kinerja terbaik dalam menyelesaikan proyek. *Activity gap* yang tinggi juga menunjukkan adanya permasalahan yang serius pada tim. Ditemukan juga bahwa tim dengan distribusi kartu baik serta memiliki frekuensi pergerakan antar kolom Kanban yang tinggi cenderung memiliki produktivitas yang lebih baik.

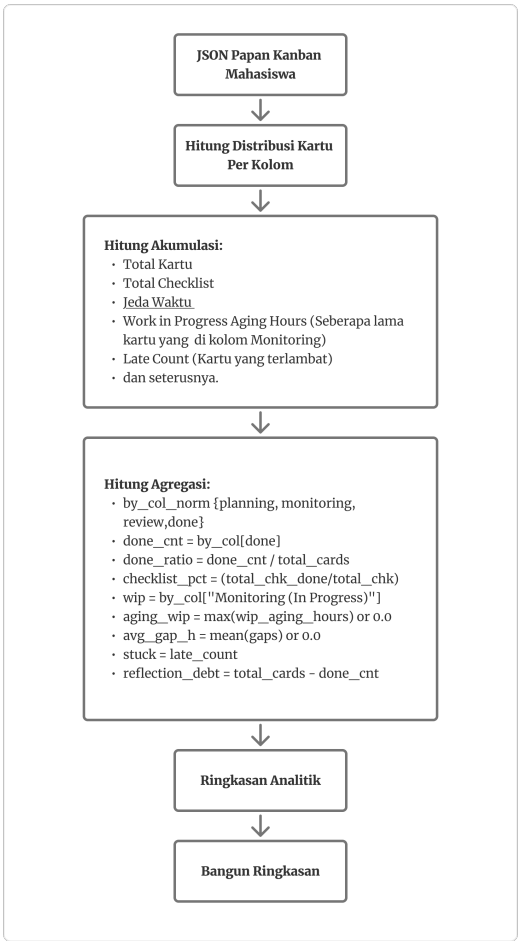
Selain itu, David J. Anderson dan Andy Carmichael (2016) [37] pada bukunya, "*Essential Kanban Condensed*" menjelaskan bahwa metrik seperti *Work In Progress* (WIP) sangat erat kaitannya dengan *Delivery Rate* (DR) atau tingkat penyelesaian tugas. WIP perlu dipantau dan dibatasi agar tidak melebihi kapasitas tim. Pembatasan WIP yang efektif dapat meningkatkan fokus dan efisiensi kerja dan memperpendek *lead-time* penyelesaian tugas. Dalam penelitian ini, indikator WIP digunakan untuk menilai beban kerja aktif mahasiswa serta memberikan umpan balik yang sesuai. Indikator WIP dimodifikasi dari membatasi jumlah kartu pada kolom Monitoring/In Progress menjadi membatasi durasi waktu kartu yang berada dalam status tersebut "`aging_thr_h`". Rata-rata jeda antar aktivitas "`avg_gap_h`" juga diadopsi untuk menilai frekuensi interaksi mahasiswa.

Dilakukan juga penetapan ambang batas untuk mendeteksi keadaan terhambat atau "`stuck`" pada kartu di WIP. Kartu dianggap terhambat jika durasi rata-rata kartu berada di WIP melebihi ambang batas "`aging_thr_h`" atau tidak ada pergerakan dalam periode tertentu. Batas tersebut ditentukan oleh variabel MOVEMENT_RECENT_HOURS untuk periode pengecekan pergerakan terbaru dan variabel AGING_WIP_THRESHOLD_H untuk ambang batas durasi kartu di WIP. Pada penelitian ini, nilai yang digunakan adalah:

- MOVEMENT_RECENT_HOURS = 2 jam
- AGING_WIP_THRESHOLD_H = 2 jam

Nilai tersebut dipilih berdasarkan batasan waktu yang sudah ditentukan pada aturan studi kasus perilaku mahasiswa yang tertera pada Tabel 3.1 di mana pergerakan kartu dalam jangka waktu 2 jam terakhir.

Dengan begitu, indikator-indikator yang akan digunakan dalam *descriptive analytics* pada penelitian ini mencakup aspek-aspek penting dari aktivitas pembelajaran mahasiswa yang dapat memberikan insight pada *Feedback Agent* untuk menghasilkan umpan balik yang lebih informatif. Diagram alur proses *descriptive analytics* yang digunakan dalam penelitian ini ditunjukkan pada Gambar 3.22.



Gambar 3.22. Proses Descriptive Analytics untuk Mendukung Umpan Balik Pedagogis

Proses *descriptive analytics* dimulai dengan menggunakan data papan Kanban mentah dalam format JSON sebagai input. Data ini kemudian diproses untuk mengekstrak data analitik seperti, menghitung distribusi kartu per kolom, total kartu pada papan, persentase *checklist* dan seterusnya. Hasil ekstraksi data analitik dilakukan agregasi dan berlanjut ke proses penyusunan ringkasan seperti pada Gambar 3.23 Setelah itu, dilakukan pembuatan ringkasan analitik mahasiswa yang akan dimasukkan ke dalam *prompt* penilaian sebagai konteks tambahan untuk *Feedback Agent* menghasilkan umpan balik pedagogis.

Berikut adalah contoh potongan ringkasan analitik mahasiswa dalam format JSON yang dihasilkan dari proses ini.

Gambar 3.23. Contoh Potongan Ringkasan Analitik Mahasiswa dalam Format JSON

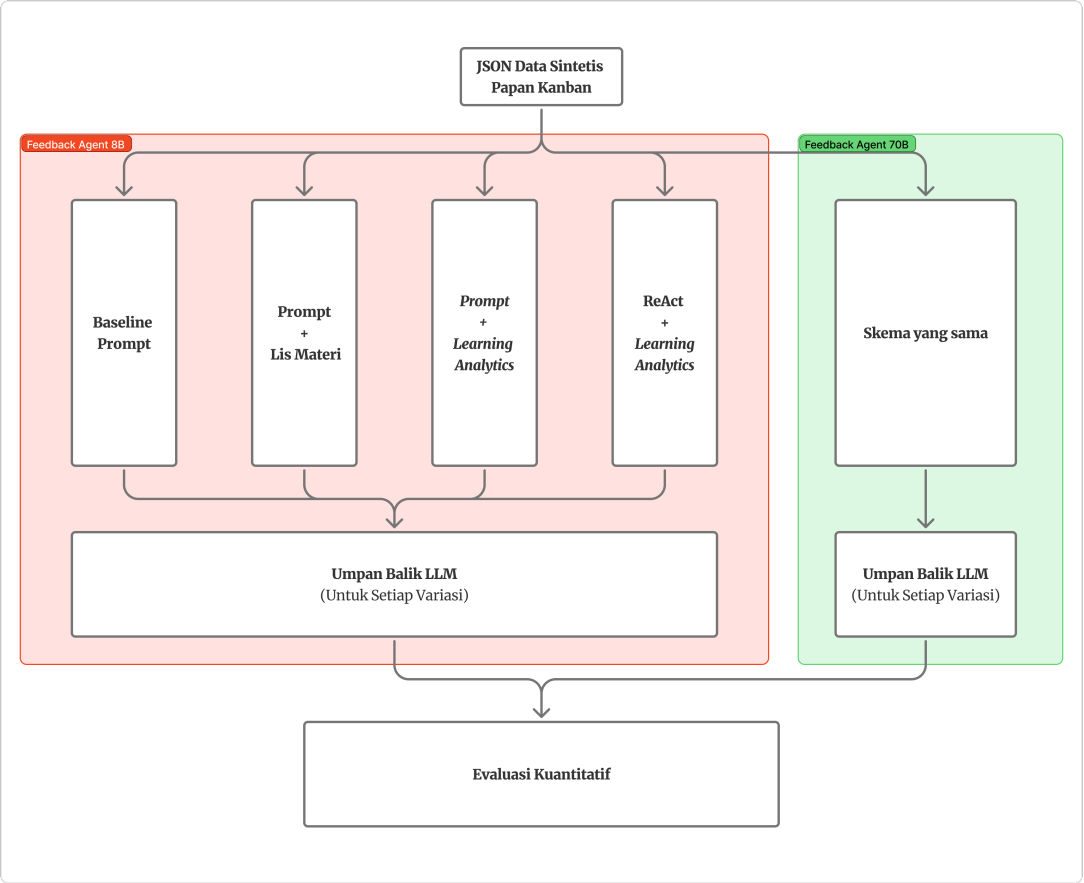
```
"Total kartu: 5; Distribusi - Planning 2, Monitoring 1, Review 0, Done 2; Checklist
selesai (agregat): 66.67%; Rasio selesai: 40%; Proses belajar terhambat (stuck): 2;
Rata-rata selisih pergerakan: 1.25 jam; Alerts: AGING\_WIP:Implementasi Pythagoras,
REFLECTION\_DEBT:3"
```

Dengan begitu, *Feedback Agent* dapat memanfaatkan data analitik mahasiswa ini untuk menghasilkan umpan balik pedagogis yang lebih terarah dan bermanfaat bagi mahasiswa.

3.3.4.3 Pembuatan Variasi *Prompt*

Pembuatan variasi *prompt* bertujuan membandingkan performa *baseline prompt* dengan *prompt* yang diberi perlakuan khusus sehingga dapat menilai dampak setiap modifikasi terhadap keluaran model. Variasi dapat meliputi penambahan informasi kontekstual berupa "Lis Materi" seperti yang tertera pada Gambar 3.19. Perlakuan lain adalah memasukkan data analitik dari *learning analytics* yang sudah dibuat sebelumnya. Selanjutnya, dibuat juga variasi mekanisme *ReAct* (*Reasoning* + *Acting*) yang digabungkan dengan data analitik. Hal ini diterapkan untuk memisahkan fase penalaran internal model dari tindakan eksternal seperti memberikan umpan balik atau menyarankan langkah berikutnya. Menggabungkan *ReAct* dengan *learning analytics* juga memungkinkan agen memberi penjelasan langkah demi langkah sekaligus rekomendasi yang relevan berdasarkan indikator pembelajaran.

Berikut skema diagram alur uji coba variasi *prompt* yang akan dilakukan dapat dilihat pada Gambar 3.24.

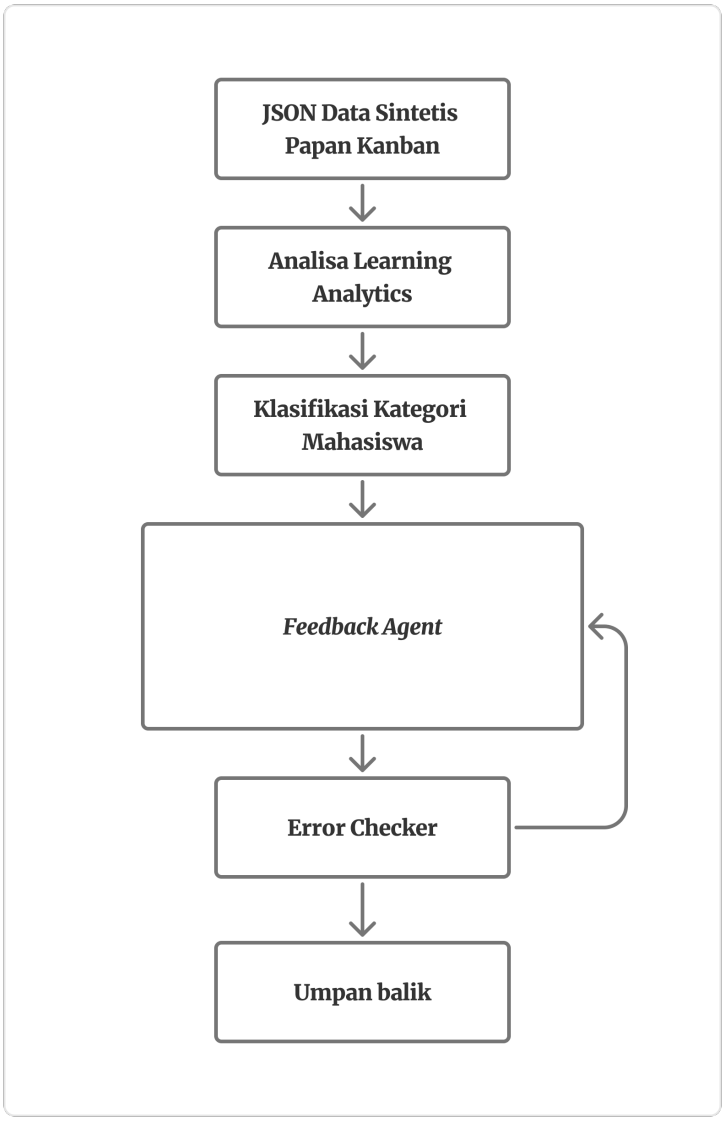


Gambar 3.24. Skema Uji Coba Variasi *Prompt*

Dari Gambar 3.24 di atas, dapat dilihat bahwa terdapat 4 variasi *prompt* yang akan diuji coba. Variasi pertama adalah *Baseline Prompt* yang tidak memiliki perlakuan khusus. Variasi kedua adalah *Prompt* dengan konteks tambahan berupa "Lis Materi" seperti yang tertera pada Gambar 3.19. Variasi ketiga adalah *Prompt* yang memasukkan data analitik dari *learning analytics* seperti yang tertera pada Gambar 3.23. Variasi keempat adalah *Prompt* dengan mekanisme *ReAct* yang digabungkan dengan data analitik dari *learning analytics*.

Hal yang sama dilakukan pada setiap ukuran model LLM yaitu Llama 3.1 8B Instant dan Llama 3.3 70B Versatile. Hasil dari uji coba variasi *prompt* ini akan disajikan pada bab hasil dan pembahasan untuk menganalisis dampak setiap perlakuan terhadap kualitas umpan balik pedagogis yang dihasilkan.

Untuk detail ReAct yang digunakan pada variasi keempat, adalah sebagai Gambar 3.25.



Gambar 3.25. Skema Mekanisme ReAct dengan Learning Analytics

ReAct *Agent* yang dibangun pada penelitian ini menggunakan *framework* dari LangGraph. Alur kerja ReAct sendiri dibuat agar model LLM dapat memiliki *checkpoint* pada setiap langkah penalaran internalnya sebelum menghasilkan umpan balik. Dengan begitu, diharapkan model dapat lebih mudah untuk melakukan refleksi dan perbaikan terhadap proses penalarannya sendiri. ReAct Agent yang dikembangkan juga dilengkapi dengan *Error Checker* yang berfungsi untuk memeriksa kesalahan pada format keluaran sehingga dapat meminimalisir kesalahan format.

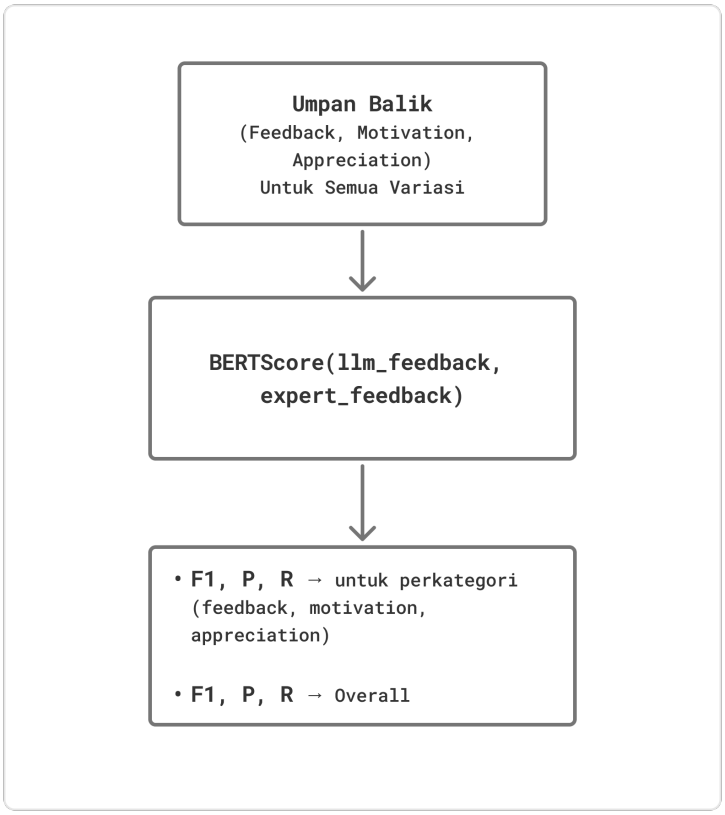
3.3.5 Alur Evaluasi Kualitas Umpan Balik Pedagogis

3.3.5.1 Evaluasi Kuantitatif

Metode evaluasi kuantitatif digunakan untuk mengukur kualitas umpan balik pedagogis yang dihasilkan oleh *Feedback Agent*. Evaluasi ini menggunakan metrik seperti F1-Score, Precision, dan Recall. Metode-metode yang digunakan seperti yang sudah dibahas pada Subbab 3.2.3 yaitu BERTScore, BARTScore, dan *LLM-as-a-Judge*. Berikut adalah penjelasan alur evaluasi kuantitatif yang digunakan dalam penelitian ini untuk tiap-tiap metodenya.

BERTScore

Alur evaluasi kuantitatif menggunakan BERTScore adalah sebagai berikut:



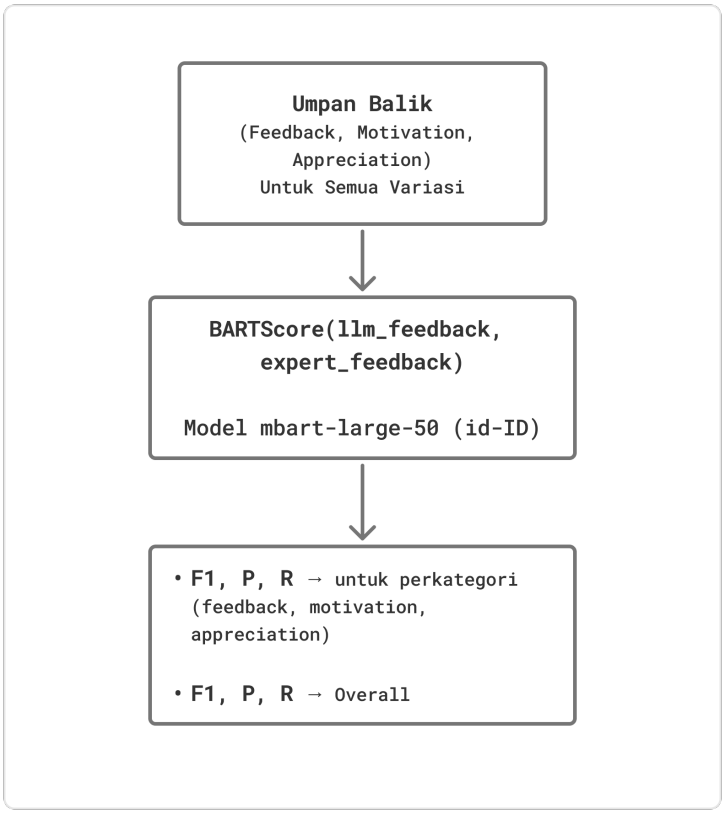
Gambar 3.26. Alur Evaluasi Kuantitatif dengan BERTScore

Dari Gambar 3.26 di atas, dapat dilihat bahwa proses evaluasi dimulai dengan mengumpulkan umpan balik pedagogis yang dihasilkan oleh *Feedback Agent* dari berbagai variasi *prompt*. Setelah itu, dilakukan persiapan data dengan mempasangkan data umpan balik yang dihasilkan dengan data ground truth dari ahli manusia. Kemudian, dilakukan perhitungan BERTScore untuk setiap pasangan umpan balik yang ada. Hasil

perhitungan BERTScore adalah skor numerik F1-Score, Precision, dan Recall dari setiap kategori umpan balik (yaitu *feedback*, *motivation*, dan *appreciation*). Selanjutnya, dihitung juga untuk skor rata-rata keseluruhan dari semua kategori untuk F1-Score, Precision, dan Recall.

BARTScore

Alur evaluasi kuantitatif menggunakan BARTScore adalah sebagai berikut:



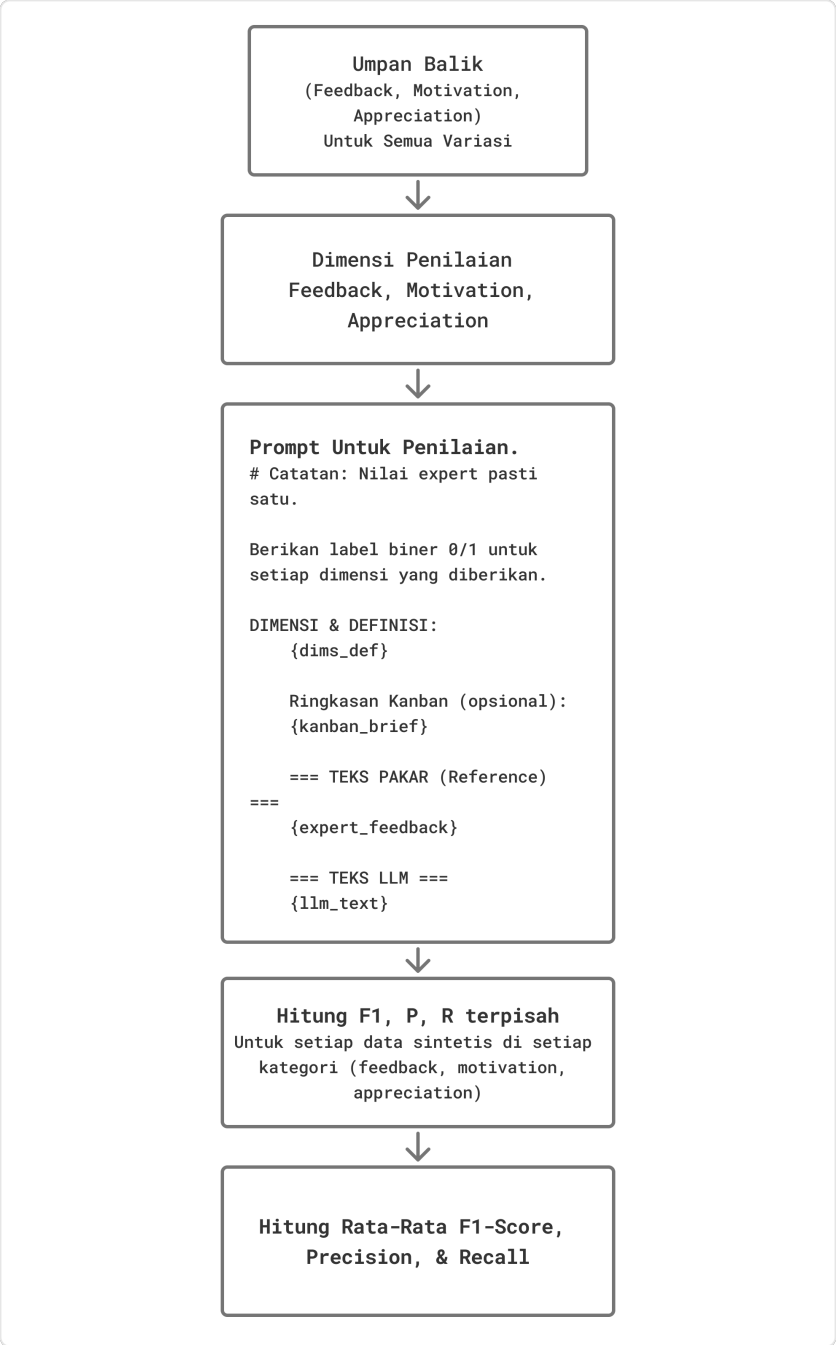
Gambar 3.27. Alur Evaluasi Kuantitatif dengan BARTScore

Dari Gambar 3.27 di atas, sama seperti BERTScore, dapat dilihat bahwa proses evaluasi dimulai dengan mengumpulkan umpan balik pedagogis yang dihasilkan oleh *Feedback Agent* dari berbagai variasi *prompt*. Setelah itu, dilakukan persiapan data dengan mempasangkan data umpan balik yang dihasilkan dengan data ground truth dari ahli manusia. Kemudian, dilakukan perhitungan BARTScore untuk setiap pasangan umpan balik yang ada. Model yang digunakan adalah model *facebook/mbart-large-50*.

Hasil perhitungan BARTScore adalah skor numerik F1-Score, Precision, dan Recall dari setiap kategori umpan balik (yaitu *feedback*, *motivation*, dan *appreciation*). Selanjutnya, dihitung juga untuk skor rata-rata keseluruhan dari semua kategori untuk F1-Score, Precision, dan Recall.

LLM-as-a-Judge

Alur evaluasi kuantitatif menggunakan *LLM-as-a-Judge* adalah sebagai berikut:



Gambar 3.28. Alur Evaluasi Kuantitatif dengan LLM-as-a-Judge

Pada proses evaluasi dengan *LLM-as-a-Judge*, digunakan model LLM GPT-4o dengan alur kerja seperti pada Gambar 3.28 di atas. Proses evaluasi dimulai dengan mengumpulkan umpan balik pedagogis yang dihasilkan oleh *Feedback Agent* dari berbagai variasi *prompt*. Lalu, pasangan data umpan balik yang dihasilkan dengan data ground

truth dari ahli manusia akan melalui proses penilaian dari model GPT-4o. Model melalui serangkaian *prompt* "Dimensi Penilaian" yang diambil dari literatur [28–31], literatur yang sama yang digunakan untuk mengevaluasi umpan balik secara kualitatif, untuk menilai kualitas umpan balik berdasarkan kriteria tertentu. Kriteria tersebut dapat dilihat pada Tabel 3.7 sebagai berikut.

Tabel 3.7. Dimensi Penilaian Umpan Balik Pedagogis

Feedback (F)	
F1	Teks jelas, mudah dipahami, tidak bertele-tele.
F2	Menyebut detail spesifik konteks/kanban (kartu, kolom, checklist, waktu/pergerakan).
F3	Sesuai dengan tugas/konteks; tidak generik.
F4	Memberi saran perbaikan proses, mendiagnosis hambatan.
F5	Memberi langkah selanjutnya (next step) yang konkret.
Motivasi (M)	
M1	Menumbuhkan motivasi intrinsik (dorongan dari dalam).
M2	Mengaitkan pada tujuan yang jelas/terukur.
M3	Mendorong otonomi/kendali diri pembelajar.
M4	Menguatkan keyakinan diri/kompetensi (self-efficacy).
M5	Menunjukkan empati/keterhubungan.
Apresiasi (A)	
A1	Mengakui usaha dan proses.
A2	Pujian selaras capaian, tidak berlebihan.
A3	Meningkatkan kepercayaan diri.
A4	Memberi dukungan emosional/menenangkan.
A5	Penguatan positif yang mendorong keberlanjutan.

Setelah melalui proses penilaian oleh model dengan melihat dimensi penilaian pada Tabel 3.7, data Sintetis papan pembelajaran Kanban, dan data ground truth dari ahli manusia. Penilaian oleh model GPT-4o memberikan nilai 0 atau 1 untuk setiap dimensi penilaian pada Tabel 3.7. Nilai 1 menunjukkan bahwa umpan balik memenuhi kriteria dan 0 menunjukkan bahwa umpan balik tidak memenuhi kriteria. Hasil tersebut akan digunakan untuk menghitung skor F1-Score, Precision, dan Recall dari setiap kategori. Lalu, diakhir penghitungan, dihitung skor rata-rata keseluruhan dari semua kategori untuk F1-Score, Precision, dan Recall.

3.3.5.2 Evaluasi Kualitatif

Metode evaluasi kualitatif digunakan untuk menilai kualitas umpan balik pedagogis yang dihasilkan oleh penilai ahli manusia bernama Shafira Anissa yang memiliki latar belakang di bidang Psikologi Pendidikan. Evaluasi ini menggunakan dimensi penilaian yang diadaptasi dari literatur [28–31]. Dimensi penilaian tersebut terdiri dari tiga kategori utama yaitu *Feedback*, *Motivation*, dan *Appreciation* dengan masing-masing kategori memiliki lima sub-dimensi penilaian. Dari literatur tersebut, dilakukan adaptasi dimensi penilaian yang menghasilkan Tabel 3.7 yang sudah ditampilkan pada Subbab sebelumnya. Namun, untuk memudahkan penilai manusia dalam melakukan evaluasi, dibuatlah kuesioner yang digunakan oleh ahli penilai. Kuesioner terdiri dari dua bagian yakni bagian penilaian skala Likert 1-5 untuk setiap sub-dimensi dan bagian komentar terbuka untuk memberikan masukan tambahan. Berikut adalah bagian pertama dari kuesioner yang digunakan untuk evaluasi secara kualitatif pada Tabel 3.8.

Tabel 3.8. Kuesioner Bagian 1 untuk Evaluasi Kualitatif

Dimensi Penilaian
Feedback (F)
F1: Kejelasan
Umpan balik dari chatbot mudah dipahami dan tidak membingungkan.
F2: Spesifisitas
Chatbot memberikan umpan balik yang spesifik terhadap kemajuan saya, bukan hanya komentar umum.
F3: Relevansi
Umpan balik yang diberikan sesuai dengan tugas atau aktivitas yang sedang saya kerjakan.
F4: Konstruktivitas
Umpan balik membantu saya memperbaiki kesalahan atau meningkatkan kinerja belajar saya.
F5: Feed-forward
Chatbot memberikan saran atau langkah selanjutnya yang dapat saya lakukan untuk memperbaiki hasil belajar.
Motivasi (M)
M1: Dorongan intrinsik
Chatbot membuat saya lebih termotivasi untuk belajar karena saya merasa mampu berkembang.
M2: Dorongan tujuan
Umpan balik dari chatbot membantu saya tetap fokus pada tujuan pembelajaran.
M3: Otonomi
Chatbot memberi saya ruang untuk mengambil keputusan sendiri dalam proses belajar.
M4: Kompetensi
Respon dari chatbot membuat saya merasa lebih kompeten dalam mengerjakan tugas.
M5: Keterhubungan
Chatbot memberikan dukungan yang terasa empatik dan memahami kondisi saya.
Apresiasi (A)
A1: Pengakuan usaha
Chatbot memberikan apresiasi terhadap usaha saya, bukan hanya hasilnya.
A2: Pujian proporsional
Pujian dari chatbot terasa tulus dan sesuai dengan pencapaian saya.
A3: Peningkatan rasa percaya diri
Ucapan apresiasi membuat saya lebih percaya diri terhadap kemampuan saya.
A4: Dukungan emosional
Chatbot memberikan respon yang menenangkan atau menyemangati ketika saya mengalami kesulitan.
A5: Penguatan positif
Chatbot menggunakan bahasa positif yang membuat saya merasa dihargai sebagai pelajar.

Lalu, berikut adalah bagian kedua dari kuesioner yang digunakan untuk evaluasi

secara kualitatif pada Tabel 3.9.

Tabel 3.9. Kuesioner Bagian 2 untuk Evaluasi Kualitatif (Isian)

1. Menurut Anda, aspek apa yang paling penting untuk dinilai dalam kualitas respon chatbot pembelajaran?
2. Apa saran Anda untuk pengembangan chatbot ini agar dapat memberikan dukungan belajar yang lebih efektif?
3. Menurut Anda, bagaimana peran chatbot seperti ini dapat berkontribusi terhadap peningkatan motivasi dan kemandirian belajar peserta didik?
Pernyataan Kesimpulan Penilaian Pakar: - Dapat digunakan tanpa revisi - Dapat digunakan dengan revisi minor - Perlu dikembangkan ulang secara menyeluruh
Catatan tambahan (jika ada):

Bagian kedua dari kuesioner ini berisi pertanyaan terbuka yang memungkinkan penilai manusia memberikan masukan lebih mendalam tentang kualitas umpan balik pedagogis yang dihasilkan oleh *Feedback Agent*. Masukan tersebut bertujuan untuk menilai kualitas respons chatbot pembelajaran, khususnya terkait kejelasan umpan balik, efektivitas dukungan belajar, serta kontribusinya terhadap motivasi dan kemandirian peserta didik. Pertanyaan ini juga bertujuan mengumpulkan masukan langsung dari pakar mengenai aspek yang perlu diperbaiki, sehingga pengembangan chatbot dapat diarahkan berdasarkan kebutuhan pedagogis nyata. Selain itu, bagian penilaian akhir memungkinkan ahli memberikan evaluasi menyeluruh tentang kelayakan chatbot, apakah sudah layak digunakan atau masih memerlukan revisi. Dengan demikian, kuesioner ini berfungsi sebagai instrumen untuk memvalidasi kualitas pedagogis chatbot sekaligus mengidentifikasi area peningkatan yang relevan.

Penilai akan disajikan data umpan balik hasil dari *Feedback Agent* dengan skema *context engineering* terbaik. Data umpan balik tersebut dipasangkan dengan data sintetis Kanban sehingga penilai dapat menilai kualitas umpan balik dengan konteks yang lengkap. Selanjutnya, data sintetis yang diberikan terlebih dahulu di *sampling* secara acak untuk tiap kategori studi kasus perilaku mahasiswa yakni sebanyak 3 data per kategorinya. Hal ini dilakukan untuk mengurangi beban kerja penilai manusia. Setelah itu, penilai manusia akan mengisi kuesioner yang sudah disediakan berdasarkan data ump-

an balik yang diberikan. Hasil dari kuesioner ini akan dianalisis secara deskriptif untuk mendapatkan wawasan mengenai kualitas umpan balik pedagogis yang dihasilkan yang akan dibahas pada Bab Hasil dan Pembahasan.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Ringkasan Pelaksanaan Eksperimen

Bab ini menyajikan dan membahas hasil dari metodologi penelitian yang diuraikan di Bab 3. Eksperimen dilakukan dengan menjalankan 40 dataset mahasiswa sintetis melalui empat variasi prompt yang berbeda yaitu:

- *Baseline Prompt* : *Prompt* dasar tanpa penambahan konteks apa pun.
- Lis Materi : penambahan daftar materi pembelajaran yang dipelajari mahasiswa dalam batasan studi kasusnya.
- LA : Penambahan analisis *learning analytics* berdasarkan data sintetis papan Kanban mahasiswa.
- ReAct + LA : Penggabungan metode *Reasoning and Acting* (ReAct) dengan analisis *learning analytics*.

Eksperimen ini dijalankan pada dua model, Llama 3.1 8B Instant dan Llama 3.3 70B Versatile, melalui Groq API. Hasil respons dari LLM kemudian dievaluasi secara kuantitatif (menggunakan BERTScore, BARTScore, dan LLM-as-a-Judge) dan kualitatif (melalui penilaian ahli manusia) untuk menjawab tiga pertanyaan penelitian.

4.2 Analisis Perbandingan Metode Peningkatan Umpan Balik (RQ 1)

Pada bagian ini, disajikan hasil eksperimen yang dilakukan dengan menggunakan empat variasi prompt yang berbeda untuk menghasilkan umpan balik pedagogis. Hasil dari umpan balik yang dihasilkan oleh masing-masing variasi prompt dianalisis menggunakan metrik kuantitatif seperti BERTScore, BARTScore, dan LLM-as-a-Judge. Hasil analisis akan dibandingkan untuk menentukan apakah ada peningkatan umpan balik dalam konteks pembelajaran daring.

4.2.1 Perbandingan Kualitas Semantik

Pada bagian ini disajikan ringkasan nilai BERTScore dan BARTScore yang diperoleh dari 40 dataset untuk setiap variasi prompt dan model. Tabel yang akan disajikan menampilkan rata-rata keseluruhan untuk masing-masing metrik.

Berikut adalah tabel hasil evaluasi dengan metode BERTScore untuk setiap variasi prompt dan model yang tertera pada tabel berikut.

Tabel 4.1. Ringkasan Hasil BERTScore per Variasi *Context-Engineering*

Variant	8B (Llama 3.1 8B)			70B (Llama 3.3 70B)		
	Precision	*Recall*	F1	*Precision*	*Recall*	F1
Baseline	0.6505	0.6719	0.6608	0.6567	0.6815	0.6687
Lis Materi	0.6500	0.6700	0.6597	0.6620	0.6805	**0.6710**
LA	0.6586	0.6733	0.6657	0.6529	0.6774	0.6648
ReAct + LA	0.6613	0.6716	**0.6663**	0.6561	0.6760	0.6658

Dari Tabel 4.1, nilai F1 BERTScore untuk model 8B berkisar antara 0,6597 hingga 0,6663, dengan variasi prompt "ReAct + LA" memberikan nilai tertinggi (0,6663). Sedangkan pada model 70B, nilai F1 BERTScore berkisar antara 0,6648 hingga 0,6710, di mana variasi "Lis Materi" memperoleh skor tertinggi (0,6710). Hal ini mengindikasikan bahwa untuk model kecil seperti 8B, penambahan konteks berupa daftar materi pembelajaran dan analisis *learning analytics* memberikan peningkatan kualitas semantik umpan balik yang dihasilkan oleh LLM. Namun, untuk model besar seperti 70B, penambahan konteks yang lebih sederhana (dalam hal ini hanya daftar materi pembelajaran) justru memberikan hasil yang lebih baik dibandingkan dengan penambahan konteks yang lebih kompleks seperti analisis *learning analytics* dan metode ReAct.

Selain hasil dari BERTScore, berikut adalah tabel hasil evaluasi dengan metode BARTScore untuk setiap variasi prompt dan model yang tertera pada tabel berikut.

Tabel 4.2. Ringkasan Hasil BARTScore per Variasi *Context-Engineering*

Variant	8B (Llama 3.1 8B)			70B (Llama 3.3 70B)		
	Precision	*Recall*	F1-Score	*Precision*	*Recall*	F1-Score
Baseline	-11.0907	-12.3656	-11.7282	-11.1677	-12.2132	-11.6905
Lis Materi	-11.1788	-12.5383	-11.8586	-11.0399	-12.1851	**-11.6125**
LA	-10.9045	-12.1800	-11.5422	-11.2543	-12.1306	-11.6924
ReAct + LA	-10.5800	-12.2345	**-11.4073**	-10.8933	-12.4587	-11.6760

Pada metrik BARTScore (Tabel 4.2), semua variasi prompt menunjukkan nilai negatif, yang merupakan karakteristik dari metrik ini dalam pengukuran kualitas teks. Model 8B menunjukkan peningkatan performa dari *Baseline* (-11,7282) ke variasi "ReAct + LA" (-11,4073), yang berarti kualitas umpan balik secara keseluruhan membaik dengan penambahan konteks dan metode ReAct. Model 70B juga menunjukkan pola serupa, meskipun variasi "Lis Materi" dan "ReAct+ LA" memiliki nilai yang cukup berdekatan,

dengan "Lis Materi" sedikit lebih baik. Untuk model kecil, metrik ini mendukung temuan bahwa penambahan konteks dan metode ReAct memberikan kontribusi positif terhadap kualitas semantik umpan balik. Namun, untuk model besar, semakin diberikan konteks yang kompleks tidak selalu menghasilkan peningkatan yang signifikan dalam kualitas semantik.

4.2.2 Perbandingan Kualitas Relevansi

Pada bagian ini disajikan ringkasan nilai LLM-as-a-Judge yang diperoleh dari 40 dataset untuk setiap variasi prompt dan model. Tabel yang disajikan menampilkan rata-rata keseluruhan untuk masing-masing metrik. Berikut adalah tabel hasil evaluasi dengan metode LLM-as-a-Judge untuk setiap variasi prompt dan model:

Tabel 4.3. Ringkasan Overall LLM-as-a-Judge per Metode

Metode	8B (Llama 3.1 8B)			70B (Llama 3.3 70B)		
	Precision	*Recall*	F1-Score	*Precision*	*Recall*	F1-Score
Baseline	1.0000	0.4567	0.5371	1.0000	0.6417	0.7271
Lis Materi	0.9333	0.4767	0.5627	1.0000	0.6600	**0.7571**
LA	1.0000	0.5017	0.5896	0.9333	0.5717	0.6399
ReAct + LA	1.0000	0.5217	**0.6121**	1.0000	0.5883	0.6715

Analisis kualitas relevansi menggunakan metode LLM-as-a-Judge memberikan gambaran yang lebih detail terkait aspek relevansi umpan balik yang dihasilkan. Dari Tabel 4.3, untuk model 8B, variasi "ReAct + LA" memberikan nilai F1-Score tertinggi sebesar 0,6121, yang menunjukkan keseimbangan terbaik antara presisi dan recall dalam menilai relevansi umpan balik. Variasi baseline memiliki nilai F1-Score terendah (0,5371), menandakan bahwa tanpa konteks tambahan, relevansi umpan balik lebih rendah. Pada model 70B, variasi "Lis Materi" memberikan nilai F1-Score tertinggi sebesar 0,7571, diikuti oleh baseline (0,7271) dan "ReAct + LA" (0,6715). Hal ini menunjukkan bahwa pada model yang lebih besar, penambahan daftar materi pembelajaran secara khusus meningkatkan relevansi umpan balik secara signifikan.

4.2.3 Analisis Evaluasi Kuantitatif

Analisis evaluasi kuantitatif pada eksperimen ini menunjukkan pola yang konsisten terkait pengaruh variasi *prompt* dan ukuran model terhadap kualitas umpan balik pedagogis yang dihasilkan. Berdasarkan metrik BERTScore dan BARTScore, model 8B (Llama 3.1 8B) menunjukkan peningkatan kualitas semantik yang paling signifikan ketika metode ReAct dan konteks *learning analytics* (ReAct + LA) diterapkan, dengan nilai

F1 BERTScore tertinggi sebesar 0,6663 dan peningkatan BARTScore dari -11,7282 (*Baseline*) menjadi -11,4073 (ReAct + LA). Hal ini mengindikasikan bahwa penambahan konteks yang kompleks dan metode ReAct mampu memperbaiki akurasi dan kualitas semantik umpan balik pada model yang lebih kecil.

Sebaliknya, pada model 70B (Llama 3.3 70B), peningkatan kualitas semantik lebih optimal dicapai dengan penambahan konteks yang lebih sederhana, yakni daftar materi pembelajaran (Lis Materi), yang menghasilkan nilai F1 BERTScore tertinggi sebesar 0,6710 dan BARTScore yang lebih baik dibandingkan variasi lain. Penambahan konteks kompleks seperti LA dan ReAct + LA tidak memberikan peningkatan signifikan, bahkan cenderung stagnan atau sedikit menurun. Ini menunjukkan bahwa model besar mungkin sudah memiliki kapasitas internal yang cukup untuk memahami konteks pembelajaran, sehingga konteks tambahan yang terlalu kompleks tidak selalu memperbaiki hasil.

Dari sisi relevansi umpan balik yang diukur menggunakan metode *LLM-as-a-Judge*, model 8B kembali menunjukkan performa terbaik dengan variasi ReAct + LA, yang menghasilkan F1-Score tertinggi 0,6121. Ini menegaskan bahwa ReAct dan analisis *learning analytics* memberikan kontribusi positif dalam meningkatkan relevansi umpan balik pada model berukuran kecil. Sedangkan pada model 70B, variasi Lis Materi mendominasi dengan F1-Score 0,7571, menandakan bahwa penambahan konteks berupa daftar materi pembelajaran secara khusus meningkatkan relevansi umpan balik secara signifikan, lebih baik dibandingkan Baseline dan ReAct + LA.

Secara keseluruhan, evaluasi kuantitatif ini mengindikasikan bahwa strategi peningkatan umpan balik pedagogis perlu disesuaikan dengan kapasitas model LLM yang digunakan. Model kecil memperoleh manfaat lebih besar dari penambahan konteks yang kompleks dan metode ReAct, sedangkan model besar lebih efektif dengan konteks yang sederhana dan fokus pada materi pembelajaran. Pendekatan ini penting untuk mengoptimalkan kualitas semantik dan relevansi umpan balik dalam aplikasi pembelajaran daring.

Dari hasil evaluasi kuantitatif menggunakan BERTScore, BARTScore, dan *LLM-as-a-Judge*, diputuskan untuk mengambil Llama 3.1 8B dengan variasi "ReAct + LA" untuk di analisis kualitatif lebih lanjut pada bagian berikutnya. Hal ini karena pada eksperimen kuantitatif, model memberikan hasil pengamatan eksperimen yang sejalan dengan seiring bertambahnya konteks yang diberikan, semakin baik hasilnya. Alasan untuk tidak diambilnya Llama 3.3 70B adalah karena pada model tersebut, semakin banyak konteks yang diberikan, hasil yang didapatkan semakin menurun sehingga perlu diketahui terlebih dahulu penyebabnya terjadinya hal tersebut. Pemilihan model yang lebih kecil juga mendukung efisiensi komputasi dan kecepatan respons sehingga lebih praktis untuk diterapkan dalam skala luas. Dengan begitu, analisis kualitatif dapat difokuskan pada model yang menunjukkan potensi terbaik dalam efisiensinya.

4.3 Analisis Pengaruh Ukuran Model (RQ 3)

Pada bagian ini, dilakukan analisis untuk menjawab *research question* ketiga (RQ 3) mengenai pengaruh ukuran model *Large Language Model* (LLM) terhadap kualitas umpan balik pedagogis. Analisis ini membandingkan performa antara model Llama 3.1 8B (model kecil) dan Llama 3.3 70B (model besar) menggunakan data kuantitatif yang telah disajikan pada Subbab 4.2.

4.3.1 Perbandingan Kapasitas Dasar (Baseline Performance)

Untuk memahami pengaruh ukuran model, langkah pertama adalah membandingkan performa dasar (*Baseline*) kedua model, yaitu ketika tidak ada *context-engineering* tambahan yang diterapkan.

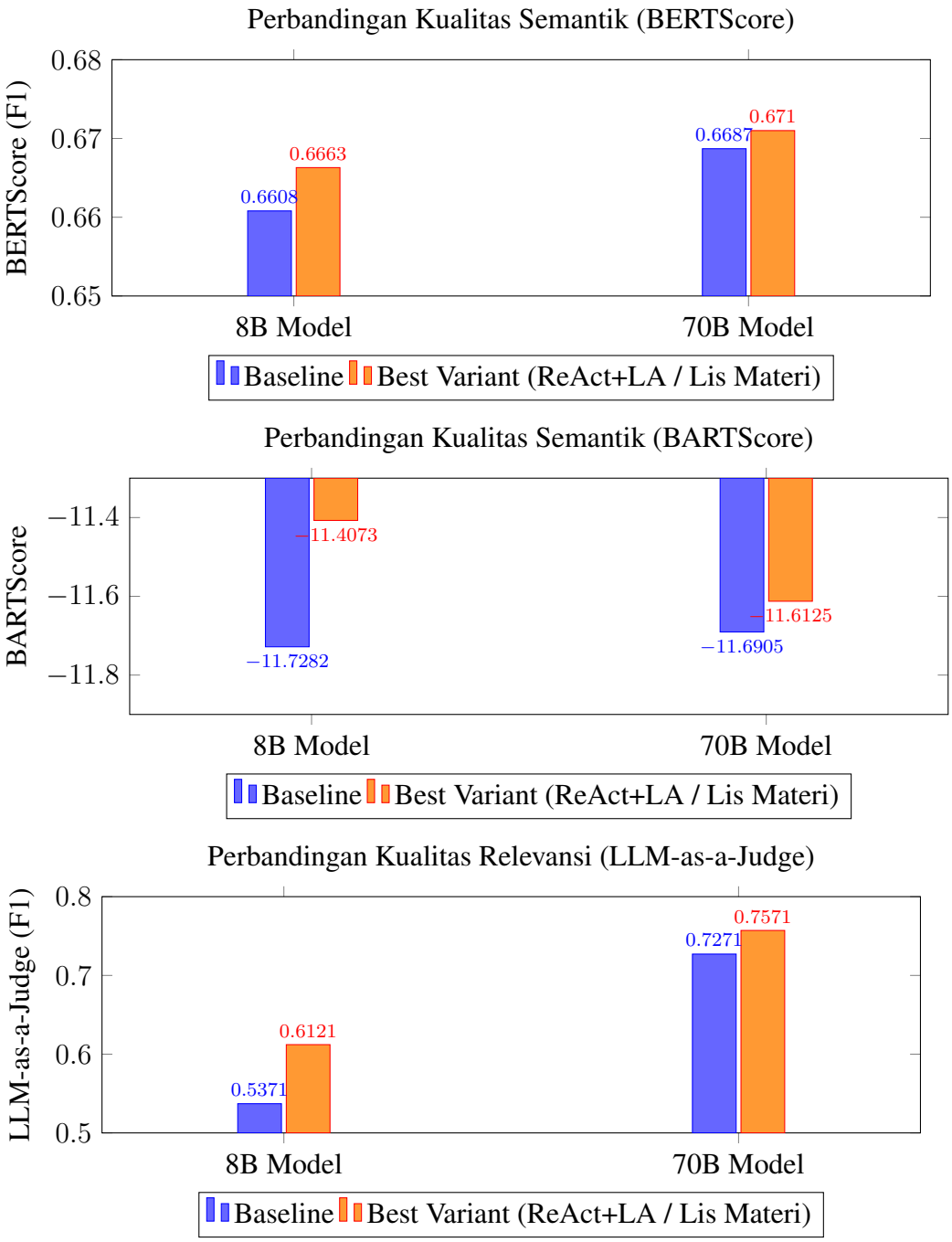
Tabel 4.4. Perbandingan Kinerja *Baseline* Antara Model 8B dan 70B

Metrik (F1-Score)	8B (Llama 3.1 8B)	70B (Llama 3.3 70B)	Pemenang
BERTScore	0.6608	0.6687	**70B**
BARTScore	-11.7282	-11.6905	**70B**
LLM-as-a-Judge	0.5371	0.7271	**70B**

Tabel 4.4 menunjukkan tren yang sangat jelas. Model 70B secara konsisten mengungguli model 8B di semua metrik evaluasi pada kondisi *Baseline*. Keunggulan paling signifikan terlihat pada metrik *LLM-as-a-Judge*, di mana model 70B (0,7271) menunjukkan skor relevansi yang jauh lebih tinggi dibandingkan model 8B (0,5371). Pada metrik semantik (BERTScore dan BARTScore), model 70B juga menunjukkan skor yang sedikit lebih baik. Hal ini mengindikasikan bahwa model yang lebih besar (70B) memiliki kapasitas internal yang lebih superior dalam menghasilkan umpan balik yang relevan dan koheren secara semantik, bahkan tanpa konteks tambahan.

4.3.2 Analisis Responsivitas terhadap *Context-Engineering*

Tren yang lebih kompleks muncul ketika menganalisis bagaimana kedua model merespons penambahan konteks (*context-engineering*). Gambar 4.29 membandingkan skor *Baseline* dengan skor terbaik yang dicapai oleh variasi *prompt* untuk masing-masing model.



Gambar 4.29. Perbandingan Kinerja Model 8B vs 70B pada metrik evaluasi berbeda.

Perbandingan kinerja antara model 8B dan 70B pada ketiga metrik evaluasi (BERTScore, BARTScore, dan *LLM-as-a-Judge*) dapat dilihat pada gambar berikut. Gambar 4.29 ini membandingkan skor F1 dari variasi *Baseline* dengan skor F1 dari variasi *prompt* terbaik untuk setiap model.

Pada metrik BERTScore, model 70B *Baseline* (0.6687) sudah lebih tinggi daripada skor terbaik model 8B (0.6663). Model 70B hanya mengalami sedikit peningkatan ke 0.6710 dengan *Lis Materi*. Sebaliknya, model 8B menunjukkan peningkatan yang lebih besar dari *Baseline* (0.6608) ke *ReAct + LA* (0.6663).

Pada metrik BARTScore, model 8B menunjukkan peningkatan paling substansial, bergerak dari -11.7282 (*Baseline*) ke -11.4073 (*ReAct + LA*), sebuah peningkatan yang signifikan. Model 70B juga meningkat, namun dengan margin yang lebih kecil, dari -11.6905 (*Baseline*) ke -11.6125 (*Lis Materi*).

Pada metrik *LLM-as-a-Judge*, kedua model menunjukkan peningkatan. Model 8B meningkat dari 0.5371 (*Baseline*) ke 0.6121 (*ReAct + LA*). Model 70B, meskipun sudah memiliki *Baseline* yang sangat tinggi (0.7271), masih dapat meningkat ke 0.7571 dengan *Lis Materi*.

Temuan kuantitatif ini, khususnya perbedaan performa yang signifikan pada metrik *LLM-as-a-Judge*, dapat dijelaskan lebih lanjut melalui analisis kualitatif terhadap luaran aktual dari kedua model dibandingkan dengan *ground truth*. Analisis kualitatif menunjukkan bahwa Llama 70B secara jelas lebih unggul daripada Llama 8B. Alasan utamanya adalah Llama 70B lebih berhasil menangkap *niat pedagogis* dari *ground truth*, sementara Llama 8B cenderung gagal dalam eksekusi dan hanya melaporkan data secara harfiah. Perbedaan mendalam ini dapat diuraikan menjadi tiga poin utama:

1. **Kualitas Feedback: Preskriptif (70B) vs. Deskriptif (8B)** Perbedaan paling signifikan terletak pada *apa* yang dilakukan oleh *feedback* tersebut.
   - ***Ground truth*:** *Feedback* bersifat preskriptif dan metakognitif. Ia tidak hanya melaporkan masalah, tetapi memberikan strategi belajar yang konkret dan dapat ditindaklanjuti.

     Contoh: "Fokus menuntaskan satu kartu sampai 'Done'...", "Hindari pindah kolom sebelum item inti selesai."
   - **Llama 3.3 70B (Baik):** Model ini *mencoba* menjadi preskriptif. Ia menganalisis data dan memberikan saran langkah berikutnya yang spesifik.

     Contoh: "Pertimbangkan untuk menyelesaikan checklist yang tersisa...", "Untuk memperbaiki ini, Anda perlu memfokuskan diri pada..."
   - **Llama 3.1 8B (Kurang):** Model ini hampir seluruhnya bersifat deskriptif. Ia hanya melaporkan data yang dilihatnya dan kemudian mengajukan pertanyaan generik, tanpa memberikan solusi atau strategi.

     Contoh: "Saya melihat bahwa Anda memiliki satu kartu di Planning...", "Saya ingin tahu apa yang menyebabkan Anda terhambat..."

   Kesimpulannya, Llama 3.3 70B bertindak sebagai pembimbing (sesuai *ground truth*), sedangkan Llama 3.1 8B hanya bertindak sebagai pelapor data.
2. **Pemisahan Fungsi: Apresiasi (70B) vs. Kritik (8B)**

Terdapat kegagalan pada Llama 3.1 8B yang berhasil diatasi oleh Llama 3.3 70B, yang juga sejalan dengan kritik ahli pada Subbab 4.4.2.

- ***Ground truth* (Target):** Apresiasinya bernilai murni, singkat, dan positif. Tujuannya hanya untuk validasi emosional.

  Contoh: "Good job, tinggal dipertahankan.", "Keberanian memulai adalah langkah penting."

- **Llama 3.3 70B (Baik):** Model ini berhasil menjaga apresiasi tetap murni. Responnya fokus pada validasi usaha dan tidak mencampurnya dengan kritik.

  Contoh: "Anda telah mengambil langkah awal yang baik...", "Ini adalah langkah yang sangat baik dan patut diapresiasi."

- **Llama 3.1 8B (Gagal):** Sementara model ini mengalami "kontaminasi respon". Ia tidak bisa memberikan apresiasi murni dan hampir selalu mencampurnya dengan *feedback* korektif, yang ditandai dengan kata "tetapi".

  Contoh: "Saya mengapresiasi usaha Anda..., **tetapi** saya ingin Anda terus maju...", "Saya mengapresiasi usaha Anda..., **tetapi** masih ada banyak yang belum diselesaikan."

Kesimpulannya, Llama 70B memahami bahwa *apresiasi* memiliki fungsi afektif yang terpisah, sementara Llama 8B gagal membedakan fungsi ini.

3. **Variasi Linguistik dan Kepatuhan Persona**

Ukuran model yang lebih besar secara langsung berdampak pada kekayaan linguistik.

- ***Ground truth* :** Memiliki variasi yang tinggi. Kata-kata motivasinya beragam: "Bangun momentum...", "Jaga ritme...", "Tetapkan target kecil...".

- **Llama 3.3 70B (Baik):** Menunjukkan variasi yang jauh lebih baik. Respon *motivasi*-nya lebih kontekstual dan tidak terlalu formulaik.

  Contoh: "Sekarang, cobalah untuk meningkatkan kompetensi Anda...", "Apa yang Anda pikirkan tentang memulai dengan menyelesaikan satu checklist lagi...?"

- **Llama 3.1 8B (Kurang):** Sangat monoton dan repetitif. Model ini terjebak pada beberapa frasa formulaik, terutama pada *motivasi*.

  Contoh: "Saya percaya bahwa Anda memiliki kemampuan untuk..." dan "Anda memiliki kemampuan untuk..." diulang di hampir setiap respon (temuan ini juga dikonfirmasi oleh ahli pada 4.4.2).

Kesimpulannya, ukuran model 70B yang lebih besar memungkinkannya memiliki

variasi linguistik yang lebih kaya, membuatnya terdengar lebih natural dan tidak robotik.

4.3.3 Interpretasi Tren dan Analisis

Setelah dilakukannya analisis komparatif antara model 8B dan 70B, dapat disimpulkan beberapa hal terkait pengaruh ukuran model terhadap kualitas umpan balik pedagogis yang dihasilkan.

1. **Kapasitas Dasar Unggul pada Model Besar**: Model 70B memiliki kapasitas dasar yang jauh lebih tinggi, terutama dalam hal relevansi (*LLM-as-a-Judge*). Skor *Baseline*-nya bahkan melampaui skor terbaik yang bisa dicapai model 8B dengan konteks paling kompleks sekalipun.
2. **Responsivitas Konteks Berbeda**: Model 8B (kecil) menunjukkan **responsivitas yang lebih tinggi** terhadap *context-engineering* yang kompleks pada pengukuran BERTScore, BARTScore, dan *LLM-as-a-Judge*. Peningkatan performa terbesar pada model 8B terjadi pada variasi "ReAct + LA", seperti yang terlihat jelas pada peningkatan skor BARTScore. Hal ini mengindikasikan bahwa model kecil sangat bergantung pada *prompt* yang terstruktur dan kaya konteks untuk mengkompensasi keterbatasan kapasitas internalnya. Hal ini juga didukung oleh literatur [41, 42] yang menyatakan bahwa model kecil cenderung lebih sensitif terhadap jumlah teknik *prompt engineering* yang digunakan dibandingkan model besar.
3. **Efisiensi Konteks pada Model Besar**: Sebaliknya, LLM 70B (besar) **tidak terlalu diuntungkan** oleh konteks yang kompleks. Performa terbaiknya justru dicapai dengan konteks yang lebih sederhana ("Lis Materi"). Penambahan konteks "ReAct + LA" yang kompleks bahkan cenderung menurunkan performanya (misalnya pada BERTScore dan *LLM-as-a-Judge*) dibandingkan dengan "Lis Materi". Hal ini juga didukung oleh literatur [41, 42] di mana model besar cenderung kurang sensitif terhadap penambahan jumlah teknik *prompt engineering* yang digunakan dibandingkan model kecil. Namun, perlu dicatat, pada literatur [41], model dengan ukuran 70B tidak disebutkan secara eksplisit, sehingga perlu penelitian lebih lanjut untuk mengonfirmasi temuan ini pada model Llama 3.3 70B. Terdapat dugaan bahwa dalam eksperimen penelitian ini, ketika *context-engineering* menjadi terlalu kompleks membuat model besar performanya "dipaksa" meniru luaran dari model Llama 3.1 8B, alih-alih mendekati *ground truth*. Terbukti dengan contoh luaran yang dihasilkan pada variasi "ReAct + LA" oleh Llama 3.3 70B yang cenderung meniru gaya bahasa Llama 3.1 8B yang deskriptif dan kurang preskriptif. Contoh luaran ini adalah seperti ini.

   - **Llama 3.3 70B (ReAct + LA):**

"Anda memiliki 1 kartu di fase Planning...", "rasio selesai masih 0%. Perlu diperhatikan juga bahwa ada alert REFLECTION_DEBT...", "Ini menunjukkan bahwa proses belajar Anda sedang terhambat."

- **Llama 3.1 8B (ReAct + LA):**

  "Saya melihat bahwa Anda memiliki satu kartu di Planning...", "Checklist selesai hanya 33,33%...", "Saya ingin tahu apa yang menyebabkan Anda terhambat..."

Luaran Llama 3.3 70B di atas menunjukkan gaya bahasa yang sangat mirip dengan Llama 3.1 8B, yang bersifat deskriptif dan kurang memberikan solusi konkret, berbeda dengan *ground truth* yang bersifat preskriptif. Hal ini mengindikasikan bahwa adanya kemungkinan penambahan konteks yang terlalu kompleks pada model besar justru memaksa model untuk meniru gaya bahasa model kecil, sehingga menurunkan kualitas umpan balik yang dihasilkannya.

Kesimpulannya, pengaruh ukuran model bersifat dua arah. Model besar (70B) memberikan kualitas dasar yang lebih baik, namun model kecil (8B) memperoleh manfaat yang jauh lebih besar dari strategi *context-engineering* yang canggih. Hal ini juga menjelaskan keputusan yang diambil pada RQ 1 di mana model 70B tidak dianalisis lebih lanjut karena penambahan konteks kompleks justru menurunkan kinerjanya, sehingga hasil tersebut yang perlu diinvestigasi terpisah dipenelitian berikutnya. Sebaliknya, model 8B dengan "ReAct + LA" dipilih karena menunjukkan pola yang logis (peningkatan performa sejalan dengan penambahan konteks) dan menawarkan efisiensi komputasi yang lebih praktis.

4.4 Validasi dan Persepsi Ahli Manusia (RQ 2)

Untuk menjawab *research question* kedua (*RQ 2*) mengenai persepsi ahli manusia terhadap kualitas respon yang dihasilkan oleh *Feedback Agent*, dilakukan proses validasi pakar. Validasi ini bertujuan untuk menilai efektivitas *Feedback Agent* sebagai asisten pembelajaran dalam memberikan *feedback*, *motivation*, dan *appreciation*. Proses ini melibatkan seorang ahli dalam bidang Psikologi Pembelajaran, Shafira Anissa, dari Universitas Gadjah Mada.

Penilaian pakar dilakukan menggunakan instrumen kuesioner terstruktur yang mencakup tiga aspek utama: *Feedback Quality* (Kualitas Umpan Balik), *Motivational Support* (Dukungan Motivasi), dan *Appreciation/Affective Support* (Apresiasi dan Penguatan Emosional). Pakar memberikan penilaian kuantitatif menggunakan skala *Likert* 1–5 serta umpan balik kualitatif melalui pertanyaan terbuka.

4.4.1 Analisis Penilaian Rubrik Ahli Manusia

Penilaian kuantitatif oleh pakar dirangkum berdasarkan 15 butir pertanyaan yang mewakili tiga aspek evaluasi. Skor 1 mewakili "Sangat Tidak Setuju", 2 "Tidak Setuju", 3 "Netral", 4 "Setuju", dan 5 "Sangat Setuju".

4.4.1.1 Aspek Kualitas Umpan Balik

Aspek ini menilai seberapa informatif, jelas, relevan, dan konstruktif umpan balik yang diberikan chatbot. Hasil penilaian pakar adalah sebagai berikut dapat dilihat pada Tabel 4.5:

Tabel 4.5. Penilaian Aspek *Feedback* Umpan Balik oleh Pakar

Indikator	Pertanyaan Angket	Skor
F1 (Kejelasan)	Feedback dari chatbot mudah saya pahami.	4
F2 (Spesifisitas)	Chatbot memberikan saran perbaikan yang spesifik terhadap pekerjaan saya.	4
F3 (Relevansi)	Feedback dari chatbot sesuai dengan aktivitas belajar yang sedang saya lakukan.	3
F4 (Konstruktivitas)	Umpan balik dari chatbot membantu saya memperbaiki kesalahan dalam belajar.	4
F5 (Feed-forward)	Chatbot memberikan arahan atau langkah selanjutnya yang dapat saya lakukan.	4

Aspek ini secara umum dinilai positif (Skor 4, "Setuju") pada mayoritas indikator, seperti Kejelasan (F1), Spesifisitas (F2), Konstruktivitas (F4), dan Feed-forward (F5). Hal ini menunjukkan *chatbot* telah berhasil menyampaikan umpan balik yang dapat dipahami dan berorientasi pada perbaikan. Namun, skor Netral (Skor 3) pada F3 (Relevansi) menjadi catatan kritis. Ini mengindikasikan bahwa meskipun umpan balik yang diberikan berkualitas baik, namun terkadang belum sepenuhnya relevan atau selaras dengan konteks aktivitas spesifik yang sedang dilakukan pengguna.

Kekurangan ini terletak pada kurangnya spesifisitas kontekstual. Sebagai contoh, respon seperti "Saya melihat bahwa Anda memiliki satu kartu di Planning..." atau "kartu yang masih dalam fase Monitoring..." tidak secara eksplisit menyebutkan *nama* kartu yang dimaksud. Hal ini membuat umpan balik terasa generik dan kurang relevan, terutama jika pengguna memiliki beberapa kartu dalam fase yang sama. Sebaliknya, respon yang lebih efektif secara spesifik menyebutkan nama kartu, seperti "'Object Oriented Programming [CS101]'". Inkonsistensi dalam menyebutkan konteks spesifik inilah yang diduga kuat menjadi penyebab skor F3 (Relevansi) dinilai Netral.

Analisis tersebut didukung dengan hasil wawancara dengan ahli. Ahli secara eksplisit mengkritik bahwa feedback "nggak spesifik" pada *timestamp* [00:02:23,946]. Contohnya adalah *chatbot* mengatakan "kamu ada lho kartu yang masih di fase planning... tapi kartu yang mana?" (*timestamp* [00:02:23,946]).

4.4.1.2 Aspek Dukungan Motivasi

Berikutnya, aspek ini menilai kemampuan chatbot dalam meningkatkan motivasi belajar. Hasil penilaian pakar dapat dilihat pada Tabel 4.6:

Tabel 4.6. Penilaian Aspek Dukungan *Motivation* oleh Pakar

Indikator	Pertanyaan Angket	Skor
M1 (Dorongan Intrinsik)	Chatbot membuat saya lebih bersemangat untuk belajar.	4
M2 (Dorongan Tujuan)	Chatbot membantu saya tetap fokus pada tujuan pembelajaran.	5
M3 (Otonomi)	Chatbot memberi saya kebebasan dalam memilih cara belajar yang sesuai.	4
M4 (Kompetensi)	Respon chatbot membuat saya merasa lebih percaya diri dalam belajar.	4
M5 (Keterhubungan)	Chatbot memberikan dukungan yang terasa empatik dan memahami kondisi saya.	3

Aspek ini menunjukkan kinerja terkuat dari chatbot, dibuktikan dengan skor Sangat Setuju (Skor 5) pada M2 (Dorongan Tujuan). Chatbot dinilai sangat efektif dalam menjaga fokus pengguna pada tujuan pembelajaran. Indikator lain seperti Dorongan Intrinsik (M1), Otonomi (M3), dan Kompetensi (M4) juga dinilai kuat (Skor 4). Namun, sama seperti aspek feedback, terdapat skor Netral (Skor 3) pada M5 (Keterhubungan). Ini menyiratkan bahwa chatbot berhasil secara fungsional dalam memotivasi, namun gagal dalam membangun koneksi empatik atau relasional dengan pengguna.

Kemungkinan penyebab skor rendah pada M5 (Keterhubungan) ini dijelaskan melalui analisis pola kalimat pada data respon motivasi. Terdapat repetisi formulaik yang sangat tinggi. Sebagian besar respon motivasi dimulai dengan frasa yang hampir identik: "Saya percaya bahwa Anda memiliki kemampuan untuk..." atau "Anda memiliki kemampuan untuk...". Meskipun tujuannya positif, pengulangan yang konstan ini membuat interaksi terasa skriptual, robotik, dan tidak tulus. Kegagalan untuk memvariasikan bahasa motivasi inilah yang menyebabkan respon gagal terasa "empatik" atau "memahami kondisi saya", sehingga kemungkinan besar menjadi alasan ahli memberikan penilaian

Netral. Hal ini juga disebutkan oleh ahli dalam wawancara di mana ahli mengeluhkan bahwa responnya "agak melaton" (monoton) dan "bentukannya itu-itu aja" pada *timestamp* [00:03:44,665].

4.4.1.3 Aspek Apresiasi dan Penguatan Emosional

Aspek ini menilai kemampuan chatbot dalam memberikan pengakuan dan dukungan emosional positif. Hasil penilaian pakar dapat dilihat pada Tabel 4.7:

Tabel 4.7. Penilaian Aspek *Appreciation* dan Penguatan Emosional oleh Pakar

Indikator	Pertanyaan Angket	Skor
A1 (Pengakuan Usaha)	Chatbot memberikan apresiasi terhadap usaha saya, bukan hanya hasilnya.	5
A2 (Pujian Proporsional)	Pujian dari chatbot terasa tulus dan sesuai dengan pencapaian saya.	3
A3 (Peningkatan Rasa Percaya Diri)	Ucapan apresiasi dari chatbot meningkatkan rasa percaya diri saya.	4
A4 (Dukungan Emosional)	Chatbot memberikan respon yang menenangkan ketika saya mengalami kesulitan.	4
A5 (Penguatan Positif)	Bahasa yang digunakan chatbot membuat saya merasa dihargai.	3

Aspek ini menunjukkan hasil yang paling beragam. Chatbot dinilai superior dalam Pengakuan Usaha (A1, Skor 5), yang sejalan dengan prinsip penguatan positif. Namun, aspek ini juga memiliki dua skor Netral (Skor 3), yaitu pada A2 (Pujian Proporsional) dan A5 (Penguatan Positif). Ini menunjukkan sebuah paradoks: chatbot memberikan apresiasi atas usaha, tetapi cara penyampaiannya (A5) dan proporsinya (A2) terasa kurang tulus atau tidak pas, sehingga membuatnya terasa kurang efektif dalam membangun rasa dihargai secara emosional.

Kemungkinan penyebab utama dari skor rendah ini adalah penggunaan "apresiasi yang dinegasikan" atau "pujian dengan 'tetapi'". Beberapa contoh respon menggunakan pola yang problematik seperti "Saya mengapresiasi usaha Anda..., **tetapi** saya ingin Anda terus maju..." atau "Saya mengapresiasi usaha Anda... **tetapi** masih ada banyak yang belum diselesaikan." Penggunaan kata 'tetapi' segera setelah pujian secara efektif membatalkan penguatan positif tersebut dan justru mengubah apresiasi menjadi *feedback* korektif. Hal ini membuat pujian terasa tidak tulus (A2) dan bahasa yang digunakan

tidak terasa menghargai (A5), yang menjelaskan penilaian Netral dari pakar. Terdapat pula tumpang tindih fungsi, di mana pesan apresiasi dicampurkan dengan *feedback*, yang mengurangi efektivitas keduanya. Dalam sesi wawancara, ahli juga secara eksplisit mengkritik penggunaan bahasa yang "negatif" seperti "utang" dan "hanya selesai" (*timestamp* [00:00:01,553]). Ahli meminta bahasa yang "lebih positif gitu ya, yang uplifting" (*timestamp* [00:00:01,553]).

4.4.2 Analisis Kualitatif Terbuka

Analisis kualitatif dilakukan terhadap jawaban pakar pada tiga pertanyaan terbuka. Umpan balik ini diuraikan berdasarkan struktur pertanyaan kuesioner untuk mendapatkan wawasan mendalam mengenai persepsi ahli terhadap kualitas, kekurangan, dan potensi kontribusi *chatbot*.

4.4.2.1 Aspek Kualitas Paling Penting

Pada pertanyaan pertama mengenai aspek paling penting dalam menilai *chatbot* pembelajaran, pakar menekankan pada dukungan substantif terhadap proses belajar. Aspek terpenting yang disoroti adalah "sejauh mana respon yang diberikan oleh *chatbot* mendukung pembelajaran peserta didik".

Kualitas ini, menurut pakar, secara spesifik dapat diukur melalui "ketepatan saran yang diberikan". Hal ini mengindikasikan bahwa fungsionalitas pedagogis dan akurasi konten (seberapa baik saran *chatbot* dalam memandu mahasiswa) dinilai sebagai prioritas utama, melampaui aspek interaksional semata.

4.4.2.2 Saran Pengembangan *Chatbot*

Pertanyaan kedua menggali saran konkret untuk pengembangan *chatbot* agar dapat memberikan dukungan yang lebih efektif. Pakar memberikan lima rekomendasi utama:

1. **Penyesuaian Tonalitas Bahasa**: Kritik utama adalah pada penggunaan bahasa yang "terlalu kaku". Pakar menyarankan agar bahasa diubah menjadi "lebih ringan" sehingga "lebih menarik dan engaging" bagi target demografi, yaitu mahasiswa. Kritik ini menyoroti adanya kontradiksi dalam desain *prompt*. Templat `INITIAL_PROMPT` menginstruksikan model untuk menggunakan nada "ramah, personal, mudah didekati". Namun, template `ASSESSMENT_PROMPT` mengubah persona tersebut menjadi "seorang dosen yang mengevaluasi" dan secara eksplisit mewajibkan keluaran dalam "Bahasa Indonesia baku (’Anda’)". Model tampaknya lebih memprioritaskan instruksi persona "dosen" yang "baku" dalam `ASSESSMENT_PROMPT`, yang secara langsung menghasilkan nada "kaku" dan mengabaikan nada "ramah" dari

INITIAL_PROMPT. Hal ini menunjukkan gagalnya model dalam mematuhi instruksi ini secara konsisten.

2. **Eliminasi Diksi Negatif**: Terkait dengan poin pertama, pakar mengidentifikasi penggunaan diksi yang terasa "negatif", seperti "utang refleksi" dan "hanya x%". Penggunaan kata-kata ini dinilai berpotensi "mengurangi motivasi peserta didik". Saran yang diberikan adalah agar *chatbot* konsisten berfokus pada "bahasa positif yang *uplifting*". Celah ini muncul bukan karena *prompt* menyuruh model menggunakan bahasa negatif, melainkan karena *prompt* tidak memiliki *guardrail* (pembatas) negatif. ASSESSMENT_PROMPT menginstruksikan model untuk memberi "komentar objektif dan konstruktif berdasarkan data" dan menggunakan "bukti dari data". Variabel di dalam <ANALISIS_JSON> berisi *string* seperti "reflection_debt" dan kalkulasi "checklist_pct" sehingga model, dalam upayanya untuk "objektif", hanya melaporkan data tersebut secara harfiah. *Prompt* tidak memiliki instruksi sekunder untuk memparafrasa data negatif menjadi kalimat yang *uplifting*. Hal ini menunjukkan perlunya penambahan *guardrail* dalam *prompt* untuk menghindari penggunaan diksi negatif secara eksplisit di penelitian selanjutnya.

3. **Tumpang Tindih Kategori Respons**: Pakar mengobservasi adanya tumpang tindih (overlap) konseptual, terutama antara feedback dengan motivasi, dan motivasi dengan apresiasi. Ditemukan bahwa bagian motivasi terkadang masih berisi elemen feedback (misalnya, "anda sebaiknya memeriksa..."). Meskipun demikian, pakar menilai ini "bukan masalah besar" karena ketiga aspek tersebut secara alami "memang saling beririsan" dalam satu kesatuan interaksi *chatbot*. Ini adalah kegagalan model dalam mematuhi *separation of concerns* (pemisahan ranah) yang ketat yang telah dirancang dalam ASSESSMENT_PROMPT. *Prompt* tersebut sudah dengan benar mendefinisikan tiga keluaran JSON yang berbeda ('feedback', 'motivasi', 'apresiasi') dengan tujuan yang berbeda. Namun, model mengalami "kebocoran konten" (*content bleeding*), di mana instruksi untuk 'feedback' ("langkah berikutnya yang konstruktif") "bocor" ke dalam *string* 'motivasi'. Ini menunjukkan bahwa meskipun struktur JSON dipaksakan, model masih kesulitan memisahkan niat pedagogis antar-kunci.

4. **Kurangnya Spesifisitas Feedback**: Poin kritik penting lainnya adalah feedback yang diberikan "belum memberikan saran yang spesifik". Pakar memberi contoh respon *chatbot* seperti, "terlihat ada kartu yang masih di fase x". Respon ini dinilai generik karena tidak menjawab "kartu yang mana yang dibicarakan?". Direkomendasikan agar *chatbot* "menunjukkan dengan lebih spesifik kartu yang sedang dibicarakan". Ini juga merupakan kegagalan eksekusi model yang paling jelas. *Prompt* INITIAL_PROMPT secara eksplisit memerintahkan model: "**Selalu rujuk data papan (nama kartu, isi checklist...)** saat memberi saran". Templat

USER_PROMPT_LA_DATA juga telah menyediakan data mentah melalui variabel <DATA>. Fakta bahwa model menghasilkan *feedback* generik "ada kartu" alih-alih mengambil nama kartu spesifik dari <DATA> menunjukkan bahwa model gagal mematuhi instruksi eksplisit untuk merujuk <DATA> demi spesifisitas.

5. **Minimnya Variasi Respon Motivasi**: Respon motivasi dinilai "kurang bervariasi" dan cenderung repetitif. Pakar mencontohkan bahwa *chatbot* terlalu sering menggunakan kalimat seperti, "saya percaya anda memiliki kemampuan untuk menyelesaikan tugas ini". Hal ini kemungkinan besar disebabkan oleh model gagal mematuhi terhadap instruksi kompleks dalam ASSESSMENT_PROMPT. *Prompt* tersebut meminta model untuk mendasarkan motivasi pada "prinsip Self-Determination Theory" (SDT). Frasa "saya percaya anda memiliki kemampuan..." adalah cara paling sederhana dan "aman" bagi model untuk memenuhi sub-prinsip "Tingkatkan Kompetensi" dari SDT. Daripada berimprovisasi, model menemukan satu frasa yang paling sesuai dengan instruksi teoretis tersebut dan menggunakannya berulang kali. Perlu diperhatikan juga bahwa *temperature* model diatur sangat rendah (0) sehingga kemungkinan besar menjadi faktor kurang bervariasinya keluaran model.

4.4.2.3 Kontribusi *Chatbot* terhadap Motivasi dan Kemandirian Belajar

Pertanyaan terakhir mengevaluasi persepsi ahli terhadap kontribusi dan peran *chatbot* dalam SRL. Ahli memandang *chatbot* ini memiliki potensi signifikan untuk berfungsi sebagai "*scaffolding* yang aksesibel" bagi peserta didik.

Keunggulan utamanya adalah aksesibilitas: *chatbot* dapat "menemani pembelajaran peserta didik kapan saja tanpa batasan waktu". Hal ini kontras dengan ketersediaan dosen atau guru yang terbatas pada waktu-waktu tertentu, di mana ahli menyatakan bahwa pengguna "Nggak usah ditungguin... harus tungguin mereka ada di kantor" (*timestamp* [00:03:44,665]). *Chatbot* dapat berfungsi "meskipun ga ada orang asli yang mendorong-dorongnya" (*timestamp* [00:00:01,553]), sehingga memberikan "*advantage*" atau "Kekuatannya" yang cukup besar. Ketersediaan 24/7 ini memungkinkan peserta didik untuk "belajar sesuai dengan *pace* masing-masing".

Pakar menyimpulkan bahwa kombinasi dari "kata-kata apresiasi, motivasi, dan *feedback*" yang dihasilkan oleh *chatbot* memiliki potensi untuk "mendorong pengguna untuk lebih semangat belajar" yang diungkapkan pada *timestamp* .

4.4.3 Validasi Kualitas oleh Ahli Manusia

Berdasarkan keseluruhan penilaian, ahli memberikan kesimpulan akhir bahwa instrumen kuesioner ini dinyatakan "**Dapat digunakan dengan revisi minor**". Revisi minor yang disarankan oleh pakar lebih berfokus pada perbaikan *chatbot* itu sendiri. Poin revisi utama meliputi penggunaan bahasa *chatbot* yang dinilai "terlalu kaku" dan kurang

sesuai dengan target demografi mahasiswa. Selain itu, pakar juga menyoroti perlunya menambah variasi dalam kalimat motivasi dan apresiasi agar tidak terkesan monoton.

Kritik penting lainnya adalah kurangnya spesifisitas *feedback*. Pakar menyarankan agar *chatbot* dapat meningkatkan kekhususan umpan balik, misalnya dengan secara eksplisit "menunjukkan... kartu yang sedang dibicarakan". Pakar menyimpulkan bahwa dengan perbaikan pada aspek aspek tersebut, *chatbot* diharapkan dapat menjadi "lebih *engaging* dan berguna" bagi peserta didik dalam mendukung proses belajar mandiri mereka.

4.5 Kelebihan dan Kekurangan Penelitian

Kelebihan utama dari penelitian ini adalah sebagai berikut.

1. Penelitian ini merupakan penelitian pertama (sejauh pengetahuan penulis) yang menggabungkan data pembelajaran mahasiswa berbasis Kanban dengan LLM untuk menghasilkan umpan balik pedagogis yang dirancang khusus mendukung proses pembelajaran mandiri (SRL).
2. Penelitian ini juga (sejauh pengetahuan penulis) merupakan penelitian pertama yang meneliti dan mengevaluasi performa LLM dalam konteks pendidikan terutama yang berfokus pada dukungan SRL mahasiswa di dalam bahasa Indonesia. Hal ini memberikan kontribusi penting terhadap literatur yang masih sangat minim di bidang ini terutama dalam bahasa Indonesia.
3. Penelitian ini berhasil menemukan hasil yang jarang ditemukan dalam literatur sebelumnya, yaitu bahwa model yang lebih besar (Llama 3.3 70B) tidak selalu lebih baik dalam menghasilkan umpan balik pedagogis jika diberikan konteks yang kompleks.
4. Penelitian ini juga berhasil menerapkan metode validasi pakar manusia untuk menilai dan menganalisa kualitas umpan balik pedagogis yang dihasilkan oleh LLM, sehingga memberikan perspektif kualitatif yang mendalam mengenai efektivitas *Feedback Agent*.
5. Penelitian ini berhasil mengembangkan metode *context-engineering* yang digabungkan dengan metode *learning analytics* dari data pembelajaran Kanban mahasiswa untuk meningkatkan performa LLM model kecil (Llama 3.3 8B) dalam menghasilkan umpan balik pedagogis. Hal tersebut merupakan bagian penelitian yang belum terjawab serta belum dikembangkan di literatur sebelumnya [20].
6. Hasil penelitian ini jika diterapkan secara nyata dapat menjadi solusi yang tergolong murah serta memberikan manfaat signifikan dalam mendukung proses pembelajaran mandiri mahasiswa, terutama dalam konteks pembelajaran jarak jauh (PJJ) yang semakin umum di era digital saat ini.

Dari semua kelebihan tersebut, terdapat beberapa kekurangan dalam penelitian ini yang perlu dicatat sebagai berikut.

1. Penelitian ini masih dalam tahap eksplorasi awal dengan jumlah sampel data yang terbatas (40 sample). Perlu dilakukannya penelitian lanjutan dengan jumlah sampel yang lebih besar, penggunaan data hasil pembelajaran mahasiswa yang asli, ahli manusia yang lebih banyak, serta variasi konteks pembelajaran yang lebih luas untuk menguji generalisasi temuan penelitian ini.
2. Penelitian ini belum dilakukan uji hipotesis statistik untuk mengukur signifikansi perbedaan performa antara variasi *prompt* dan hanya melihat tren performa berdasarkan metrik BERTScore dan BARTScore.
3. Penelitian ini hanya menguji dua model LLM (Llama 3.3 8B dan 70B) dikarenakan keterbatasan sumber daya dan biaya. Pengujian pada model lain, termasuk model komersial seperti GPT-5 atau GPT-OSS dapat memberikan wawasan yang lebih komprehensif mengenai performa LLM dalam konteks ini.
4. Penelitian ini belum dapat menguji performa *Feedback Agent* dalam konteks penggunaan nyata, sehingga akan diujikan pada penelitian selanjutnya.
5. Masih banyak faktor yang dapat diperbaiki dalam penelitian ini seperti substansi *prompt* yang digunakan, metode *context-engineering*, serta teknik *learning analytics* yang diterapkan.
6. Penelitian ini juga belum dapat menjawab secara pasti mengapa model yang besar (70B) tidak selalu lebih baik dibandingkan model yang lebih kecil (8B) dalam konteks penelitian ini. Hal ini memerlukan analisis yang lebih mendalam mengenai perilaku model LLM dalam konteks pembelajaran

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini telah dilakukan eksperimen untuk merancang, meningkatkan, dan mengevaluasi kualitas umpan balik pedagogis yang dihasilkan oleh LLM. Fokus utama penelitian adalah pada peningkatan tiga fungsi pedagogis yaitu *feedback*, *motivation*, dan *appreciation* yang untuk mendukung proses SRL mahasiswa. Peningkatan ini dicapai melalui metode *context engineering* yang diperkaya dengan *learning analytics* (LA) yang diekstraksi dari data sintetis pembelajaran Mahasiswa berbasis papan Kanban. Eksperimen dilakukan, tanpa melakukan pelatihan ulang (*fine-tuning*) pada model. Empat variasi *prompt* (*Baseline*, Lis Materi, LA, dan ReAct + LA) diujikan pada dua model LLM dengan ukuran berbeda (Llama 3.1 8B dan Llama 3.3 70B). Evaluasi dilakukan secara kuantitatif menggunakan metrik BERTScore, BARTScore, dan *LLM-as-a-Judge* terhadap data umpan balik ahli (*ground truth*) , serta secara kualitatif melalui validasi pakar psikologi pendidikan. Berdasarkan hasil analisis dan pembahasan pada Bab 4, berikut adalah kesimpulan yang ditarik untuk menjawab tiga rumusan masalah penelitian:

1. **(Menjawab RQ 1)** Metode *context engineering* yang digabungkan dengan *learning analytics* (LA) terbukti **berhasil meningkatkan kualitas umpan balik LLM** dibandingkan dengan *Baseline*, namun efektivitasnya sangat bergantung pada ukuran model. Pada model kecil (Llama 3.1 8B), penambahan konteks yang kompleks dan metode ReAct dan *learning analytics* ("ReAct + LA") secara konsisten memberikan peningkatan performa tertinggi di ketiga metrik evaluasi kuantitatif. Hal ini menunjukkan bahwa model kecil sangat diuntungkan oleh panduan konteks yang kaya dan terstruktur untuk menghasilkan umpan balik yang lebih relevan dan berkualitas secara semantik.
2. **(Menjawab RQ 3)** Ukuran model **memiliki pengaruh yang signifikan dan bersifat dua arah** terhadap kualitas umpan balik. Model besar (Llama 3.3 70B) menunjukkan performa *Baseline* yang jauh lebih superior dibandingkan model 8B, terutama dalam hal relevansi (skor F1 *LLM-as-a-Judge* 0,7271 vs 0,5371). Namun, model 70B tidak diuntungkan oleh konteks yang kompleks di mana performa terbaiknya justru dicapai dengan konteks sederhana ("Lis Materi"). Penambahan konteks "ReAct + LA" yang kompleks bahkan menurunkan kinerjanya, diduga karena *prompt* tersebut terlalu membatasi dan memaksa model besar meniru gaya deskriptif model kecil, alih-alih mendekati jawaban dari *ground truth*.
3. **(Menjawab RQ 2)** Berdasarkan persepsi ahli manusia, umpan balik yang dihasilkan oleh model terbaik (Llama 3.1 8B dengan "ReAct + LA") dinyatakan **"Dapat digu-**

nakan dengan revisi minor". Pakar mengonfirmasi potensi *chatbot* sebagai sistem *scaffolding* yang aksesibel. Revisi minor yang disarankan berfokus pada penyesuaian tonalitas bahasa, eliminasi diksi negatif, tumpang tindihnya kategori respons, kurangnya spesifitas feedback, dan minimnya variasi respons pada motivasi. Perlu dicatat bahwa beberapa instruksi dari *prompt* juga turut andil dalam beberapa revisi minor yang disarankan. Hal ini juga menegaskan pada penilaian kualitas *prompt* yang digunakan berdasarkan penelitian dari [17] yaitu *theory-driven prompt manual*, di mana nilai yang baik dari sebuah *prompt* belum tentu menjamin keluaran yang optimal dari model. Faktor-faktor lain seperti ukuran model, kapabilitas bahasa, dan kompleksitas konteks juga berperan penting dalam menentukan kualitas keluaran. Hal tersebut perlu diteliti lebih lanjut pada penelitian selanjutnya.

Kesimpulan yang dihasilkan, penelitian ini menyimpulkan bahwa *context engineering* yang diperkaya *learning analytics* adalah strategi yang efektif, praktis, dan efisien secara komputasi untuk meningkatkan umpan balik pedagogis, terutama pada model LLM berukuran kecil.

5.2 Saran

Dengan melihat berbagai kekurangan yang ada serta kemungkinan potensi pengembangan yang bisa dilakukan pada penelitian kali ini, penelitian selanjutnya dapat dipertimbangkan berbagai pengembangan sebagai berikut.

1. Meneliti lebih lanjut mengenai pengaruh *context engineering* yang kompleks pada LLM berukuran besar, untuk memahami batasan dan karakteristiknya dalam merespons konteks yang kaya.
2. Mengembangkan evaluasi ahli manusia dengan dengan menggunakan A/B testing, di mana ahli menilai dua variasi umpan balik yang dihasilkan sehingga dapat membandingkan serta memilih variasi yang terbaik secara langsung.
3. Melakukan eksperimen dengan variasi *prompt* yang lebih beragam, termasuk variasi yang menyesuaikan tonalitas bahasa sesuai karakteristik demografis pengguna.
4. Meneliti lebih lanjut *prompt* yang optimal untuk menghasilkan umpan balik yang berkualitas, tidak hanya berdasarkan penilaian dari *theory-driven prompt manual*.
5. Melakukan uji hipotesis statistik untuk memperkuat validitas temuan kuantitatif, serta mengembangkan arah penelitian yang sekarang berfokus pada eksploratif menjadi hipotesis yang kuat.
6. Melakukan evaluasi terhadap LLM yang memiliki performa bahasa Indonesia yang lebih baik, untuk melihat dampaknya terhadap kualitas umpan balik.
7. Menggunakan data pembelajaran nyata dari mahasiswa untuk mengekstraksi *lear-*

ning analytics, serta jumlah ahli yang lebih banyak untuk validasi pakar dan pengambilan *ground truth*, guna meningkatkan keakuratan dan relevansi hasil penelitian.

DAFTAR PUSTAKA

[1] J. Woodrow, S. Koyejo, and C. Piech, "Improving generative ai student feedback: Direct preference optimization with teachers in the loop," in *Proceedings of the 2025 International Conference on Educational Data Mining (EDM 2025)*. International Educational Data Mining Society, 2025, p. —. [Online]. Available: https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.short-papers.166/index.html

[2] B. Zimmerman and A. Moylan, "Self-regulation: Where metacognition and motivation intersect," *Handbook of metacognition in education*, pp. 299–315, 01 2009.

[3] C. Liu, Z. A. Bakar, and Q. Xu, "Self-regulated learning and academic achievement in higher education: A decade systematic review," *International Journal of Research and Innovation in Social Science (IJRISS)*, vol. 9, no. 03, pp. 4488–4504, 2025. [Online]. Available: https://dx.doi.org/10.47772/IJRISS.2025.90300358

[4] S. Xiao, K. Yao, and T. Wang, "The relationships of self-regulated learning and academic achievement in university students," *SHS Web of Conferences*, vol. 60, p. 01003, 2019, pHECSS2018. [Online]. Available: https://doi.org/10.1051/shsconf/20196001003

[5] E. Chitra, N. Hidayah, M. Chandratilake, and V. D. Nadarajah, "Self-regulated learning practice of undergraduate students in health professions programs," *Frontiers in Medicine*, vol. 9, p. 803069, 2022. [Online]. Available: https://doi.org/10.3389/fmed.2022.803069

[6] A. Akdeniz, "Exploring the impact of self-regulated learning intervention on students' strategy use and performance in a design studio course," *International Journal of Technology and Design Education*, pp. 1–35, 2022, advance online publication. [Online]. Available: https://doi.org/10.1007/s10798-022-09798-3

[7] B. J. Zimmerman and A. R. Moylan, "Self-regulation: Where metacognition and motivation intersect," in *Handbook of Metacognition in Education*, D. J. Hacker, J. Dunlosky, and A. C. Graesser, Eds. New York, NY: Routledge, 2009, pp. 299–315.

[8] F. Nyman, ""you're not learning skills—you're just realizing what you can do": a preliminary study of self-regulation in higher education," *Frontiers in Education*, vol. Volume 9 - 2024, 2024. [Online]. Available: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2024.1418297

[9] L. F. Yang, Y. Liu, and Z. Xu, "Examining the effects of self-regulated learning-based teacher feedback on english-as-a-foreign-language learners' self-regulated writing strategies and writing performance," *Frontiers in Psychology*, vol. Volume 13 - 2022, 2022. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1027266

[10] T. Bock, E. Thomm, J. Bauer, and B. Gold, "Fostering student teachers' research-based knowledge of effective feedback," *European Journal of Teacher*

Education, vol. 47, no. 2, pp. 389–407, 2024. [Online]. Available: https://doi.org/10.1080/02619768.2024.2338841

[11] H. E. Schaffer, K. R. Young, E. W. Ligon, and D. D. Chapman, "Automating individualized formative feedback in large classes based on a directed concept graph," *Frontiers in Psychology*, vol. 8, p. 260, 2017, article 260. [Online]. Available: https://doi.org/10.3389/fpsyg.2017.00260

[12] A. Pardo, J. Jovanović, S. Dawson, D. Gašević, and N. Mirriahi, "Using learning analytics to scale the provision of personalised feedback," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 128–138, Jan. 2019. [Online]. Available: https://doi.org/10.1111/bjet.12592

[13] M. Usher, "Generative ai vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education," *Assessment & Evaluation in Higher Education*, 2025, published online: Apr. 9, 2025. [Online]. Available: https://doi.org/10.1080/02602938.2025.2487495

[14] K. K. Maurya, K. V. A. Srivatsa, K. Petukhova, and E. Kochmar, "Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors," *arXiv preprint arXiv:2412.09416*, 2025, version 2. [Online]. Available: https://arxiv.org/abs/2412.09416

[15] S. Strickroth, M. Kreidenweis, and Z. Wurm, "Learning from agile methods: Using a kanban board for classroom orchestration," in *Proceedings of the 25th International Conference on Interactive Collaborative Learning (ICL 2022)*, ser. Lecture Notes in Networks and Systems, M. E. Auer, W. Pachatz, and T. Rüütmann, Eds., vol. 633. Springer, 2022, pp. 68–79. [Online]. Available: https://doi.org/10.1007/978-3-031-26876-2_7

[16] A. Scarlatos, D. Smith, S. Woodhead, and A. Lan, "Improving the validity of automatically generated feedback via reinforcement learning," in *Artificial Intelligence in Education*, A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds. Cham: Springer Nature Switzerland, 2024, pp. 280–294.

[17] L. J. Jacobsen and K. E. Weber, "The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of ai-driven feedback," *AI*, vol. 6, no. 2, 2025. [Online]. Available: https://doi.org/10.3390/ai6020035

[18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022. [Online]. Available: https://arxiv.org/abs/2203.02155

[19] H. Li and B. Ma, "Design of ai-powered tool for self-regulation support in programming education," 2025. [Online]. Available: https://arxiv.org/abs/2504.03068

[20] S. Strickroth, M. Kreidenweis, and A. Götzfried, "Supporting agile classroom orchestration with a live teacher dashboard," in *Futureproofing Engineering Education for Global Responsibility (ICL 2024)*, ser. Lecture Notes in Networks and Systems, vol. 1260. Cham: Springer, 2025, pp. 38–50.

[21] L. Tang, Z. Sun, B. Idnay, J. G. Nestor, A. Soroush, P. A. Elias, Z. Xu, Y. Ding, G. Durrett, J. F. Rousseau, C. Weng, and Y. Peng, "Evaluating large language models on medical evidence summarization," *npj Digital Medicine*, vol. 6, no. 1, p. 158, Aug. 2023. [Online]. Available: https://www.nature.com/articles/s41746-023-00896-7

[22] R. Lakatos *et al.*, "Investigating the performance of retrieval-augmented generation and fine-tuning for the development of ai-driven knowledge-based systems," 2024. [Online]. Available: https://arxiv.org/pdf/2403.09727

[23] E. Sulem, O. Abend, and A. Rappoport, "Bleu is not suitable for the evaluation of text simplification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 738–744. [Online]. Available: https://aclanthology.org/D18-1081

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations (ICLR)*, 2020, also available as arXiv:1904.09675.

[25] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 27 263–27 277.

[26] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using gpt-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153/

[27] C. Koutchéme, N. Dainese, S. Sarsa, A. Hellas, J. Leinonen, and P. Denny, "Open source language models can provide feedback: Evaluating llms' ability to help students using gpt-4-as-a-judge," 2024.

[28] V. J. Shute, "Focus on formative feedback," *Review of Educational Research*, vol. 78, no. 1, pp. 153–189, 2008.

[29] J. Hattie and H. Timperley, "The power of feedback," *Review of Educational Research*, vol. 77, no. 1, pp. 81–112, 2007.

[30] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum, 1985.

[31] B. L. Fredrickson, "The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions," *American Psychologist*, vol. 56, no. 3, pp. 218–226, 2001.

[32] S. Paris, J. Byrnes, and A. Paris, "Constructing theories, identities, and actions of self-regulated learners," *Self-regulated learning and academic achievement: Theoretical perspectives*, pp. 253–287, 01 2001.

[33] B. Zimmerman, "Self-regulated learning and academic achievement: An overview," *Educational Psychologist - EDUC PSYCHOL*, vol. 25, pp. 3–17, 01 1990.

[34] K. Lobos, R. Cobo-Rendón, D. Bruna Jofre, and J. Santana, "New challenges for higher education: self-regulated learning in blended learning contexts," *Frontiers in Education*, vol. 9, p. 1457367, 2024.

[35] B. Sindhu, R. P. Prathamesh, M. B. Sameera, and S. KumaraSwamy, "The evolution of large language model: Models, applications and challenges," in *2024 International Conference on Current Trends in Advanced Computing (ICCTAC)*, Bengaluru, 2024.

[36] M. I. Azmi, "Pengembangan kanban board berbasis web untuk mendukung self-regulated learning," Master's thesis, Universitas Gadjah Mada, Yogyakarta, Indonesia, 2025, dibimbing oleh Dr. Indriana Hidayah, S.T., M.T. dan Teguh Bharata Adji, S.T., M.T., M.Eng., Ph.D. [Online]. Available: http://etd.repository.ugm.ac.id/

[37] D. J. Anderson and A. Carmichael, *Essential Kanban Condensed*, 1st ed. Lean Kanban University Press, 2016, first edition, digital version, 17 April, 2016.

[38] G. Pisoni and M. Hoogeboom, "Investigating effective dynamics of virtual student teams through analysis of trello boards," 2020, afiliasi Penulis: Galena Pisoni (Department of Information Engineering and Computer Science, University of Trento), Marcella Hoogeboom (Educational Sciences, University of Twente). Informasi penerbit, volume, dan halaman tidak tersedia dari teks yang diberikan.

[39] S. Steinert, K. E. Avila, S. Ruzika, J. Kuhn, and S. Küchemann, "Harnessing large language models to develop research-based learning assistants for formative feedback," *Smart Learning Environments*, vol. 11, no. 1, p. 62, 2024. [Online]. Available: https://slejournal.springeropen.com/articles/10.1186/s40561-024-00354-1

[40] "Groq API Pricing," Groq. [Online]. Available: https://groq.com/pricing

[41] Y. Tang, D. Tuncel, C. Koerner, and T. Runkler, "The Few-shot Dilemma: Overprompting Large Language Models," *arXiv preprint arXiv:2509.13196*, Sep 2025.

[42] L. Pawlik, "How the choice of llm and prompt engineering affects chatbot effectiveness," *Electronics*, vol. 14, no. 5, 2 2025.