EKSPERIMEN MULTI-MODEL DAN MULTI-PERSONA UNTUK MENGANALISIS DAMPAK *PERSONA* TERHADAP PENALARAN, PERILAKU KELUARAN, DAN *HUMAN BIAS* PADA LARGE LANGUAGE MODEL

Proposal Tugas Akhir

Oleh

Abel Apriliani
18222008

PROGRAM STUDI SISTEM DAN TEKNOLOGI INFORMASI
SEKOLAH TEKNIK ELEKTRO DAN INFORMATIKA
INSTITUT TEKNOLOGI BANDUNG

Desember 2025

LEMBAR PENGESAHAN

EKSPERIMEN MULTI-MODEL DAN MULTI-PERSONA UNTUK MENGANALISIS DAMPAK *PERSONA* TERHADAP PENALARAN, PERILAKU KELUARAN, DAN *HUMAN BIAS* PADA LARGE LANGUAGE MODEL

Proposal Tugas Akhir

Oleh

Abel Apriliani
18222008

Program Studi Sistem dan Teknologi Informasi
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung

Proposal Tugas Akhir ini telah disetujui dan disahkan
di Bandung, pada tanggal 2 Desember 2025

Pembimbing 1	Pembimbing 2
Dr. Eng. Ayu Purwarianti, S.T., M.T.	Dr. Alham Fikri Aji, S.T., M.Sc.
NIP. x	NIP. x

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE

BAB I

PENDAHULUAN

I.1 Latar Belakang

Kemajuan dalam pengembangan *large language model* dalam beberapa tahun terakhir telah mengubah cara sistem komputasi memahami, memproses, dan menghasilkan bahasa alami. Model seperti GPT, LLaMA, Mistral, dan Gemini dilatih menggunakan korpus dalam skala masif dan mampu menyelesaikan berbagai tugas mulai dari penalaran numerik hingga interpretasi skenario sosial. Dalam banyak kasus, model menunjukkan kemampuan yang mendekati atau bahkan melampaui performa manusia pada benchmark tertentu. Walaupun demikian, peningkatan kapabilitas ini tidak sepenuhnya diikuti oleh stabilitas perilaku model dalam konteks interaksi dunia nyata.

Salah satu fenomena yang semakin banyak diamati dalam penelitian mutakhir adalah bahwa perilaku *large language model* tidak hanya dipengaruhi oleh isi instruksi, tetapi juga oleh identitas pengguna yang tersirat atau dinyatakan secara eksplisit dalam konteks percakapan. Studi mengenai bias penalaran implisit menunjukkan bahwa perubahan kecil pada deskripsi identitas pengguna dapat menyebabkan variasi signifikan pada hasil penalaran, bahkan untuk tugas yang tidak memiliki aspek sosial eksplisit (Gupta dkk. 2024). Variasi ini mencakup perubahan langkah penalaran, perbedaan tingkat kehati-hatian, hingga munculnya bias tertentu terhadap kelompok sosial.

Selain *user persona* eksplisit yang dituliskan secara langsung dalam instruksi, penelitian menunjukkan bahwa model juga sensitif terhadap *user persona* implisit yang muncul melalui gaya bahasa, framing naratif, struktur pertanyaan, atau atribut linguistik lainnya (Tseng dkk. 2024). Dalam kondisi tersebut, model tidak menerima instruksi tentang identitas pengguna, tetapi tetap membentuk asumsi internal mengenai siapa pengguna dan menyesuaikan respons sesuai asumsi tersebut. Sensitivitas

ini menandakan bahwa model melakukan inferensi identitas pengguna berdasarkan sinyal linguistik yang tampak sepele, yang berimplikasi pada stabilitas penalaran dan keadilan respons.

Penelitian pada bidang pemodelan pengguna menunjukkan bahwa variasi identitas pengguna—seperti usia, latar belakang profesional, afiliasi budaya, atau posisi sosial—dapat memengaruhi keluaran model dalam berbagai dimensi, termasuk penalaran, preferensi jawaban, dan konsistensi respons (Naous, Roziere, dkk. 2025). Hal ini menunjukkan bahwa identitas pengguna, baik eksplisit maupun implisit, berfungsi sebagai variabel laten yang memengaruhi proses generatif model. Dengan demikian, analisis terhadap *user persona* menjadi penting tidak hanya untuk memahami perilaku model, tetapi juga untuk mengidentifikasi potensi bias dan ketidakstabilan yang muncul dalam interaksi manusia–AI.

Walaupun berbagai studi sebelumnya memberikan indikasi bahwa identitas pengguna memengaruhi perilaku model, penelitian yang ada masih memiliki batasan. Mayoritas studi hanya mengevaluasi satu atau dua model, cakupan persona yang terbatas, atau jenis tugas yang sempit. Selain itu, tidak banyak studi yang secara sistematis membandingkan efek *user persona* eksplisit dan implisit pada berbagai model dan berbagai jenis tugas penalaran dalam satu kerangka eksperimen yang konsisten. Belum tersedia pula pendekatan evaluasi yang secara terpadu menguji sensitivitas model terhadap variasi identitas pengguna di berbagai kondisi tugas, baik numerik, logis, faktual, sosial, maupun moral.

Kekosongan penelitian ini penting untuk dijembatani, mengingat model bahasa semakin banyak digunakan pada skenario yang sensitif terhadap identitas pengguna, seperti layanan kesehatan, pendidikan, konseling, sistem rekomendasi, dan interaksi berbasis nilai. Ketidakstabilan respons akibat identitas pengguna berpotensi menimbulkan bias, mengurangi keandalan model, dan menghasilkan ketidaksetaraan dalam pengalaman pengguna. Oleh karena itu, diperlukan pendekatan evaluasi yang lebih komprehensif untuk memahami bagaimana *user persona* eksplisit dan implisit memengaruhi penalaran, perilaku keluaran, dan kecenderungan *human bias* pada berbagai *large language model*.

Berdasarkan urgensi tersebut, penelitian ini disusun untuk melakukan evaluasi empiris terhadap pengaruh *user persona* eksplisit dan *user persona* implisit melalui eksperimen terstruktur pada berbagai model dan berbagai jenis tugas. Penelitian ini diharapkan memberikan pemahaman yang lebih mendalam mengenai sensitivitas model terhadap identitas pengguna serta implikasinya terhadap penalaran, bias, dan

keandalan model dalam aplikasi dunia nyata.

I.2 Rumusan Masalah

Rumusan masalah berikut disusun berdasarkan kebutuhan untuk memahami bagaimana *user persona* memengaruhi perilaku dan penalaran model bahasa. Penelitian sebelumnya menunjukkan bahwa identitas pengguna, baik yang diberikan secara eksplisit maupun implisit, dapat memengaruhi penalaran, kualitas keluaran, dan kecenderungan bias model (Gupta dkk. 2024; Tseng dkk. 2024; Naous, Roziere, dkk. 2025). Namun, cakupan penelitian terdahulu masih terbatas pada sedikit model, sedikit persona, dan variasi tugas yang sempit.

Berdasarkan kondisi tersebut, rumusan masalah penelitian ini adalah sebagai berikut.

1. Bagaimana pengaruh *user persona* eksplisit dan *user persona* implisit terhadap performa penalaran pada berbagai jenis tugas pada sejumlah *large language model*.
2. Bagaimana kedua jenis *user persona* tersebut memengaruhi perilaku keluaran model pada skenario interaksi yang berbeda.
3. Bagaimana pola *human bias* muncul dan berubah sebagai akibat variasi *user persona*.
4. Sejauh mana sensitivitas terhadap *user persona* berbeda pada berbagai *large language model*, serta model mana yang menunjukkan tingkat *robustness* yang lebih tinggi terhadap variasi tersebut.

I.3 Tujuan Penelitian

Tujuan penelitian ditetapkan untuk menjawab permasalahan yang telah dirumuskan. Penelitian ini diarahkan untuk menghasilkan pemahaman yang lebih komprehensif mengenai pengaruh *user persona* terhadap perilaku model bahasa dalam tugas penalaran dan skenario percakapan. Secara khusus, penelitian ini bertujuan untuk:

1. Menganalisis pengaruh *user persona* eksplisit dan *user persona* implisit terhadap performa penalaran pada sejumlah *large language model*.
2. Mengidentifikasi perubahan perilaku keluaran model yang diinduksi oleh variasi *user persona* pada berbagai konteks.
3. Menganalisis pola *human bias* yang muncul akibat variasi *user persona*.
4. Menyusun perbandingan sensitivitas dan *robustness* berbagai model terhadap variasi *user persona*.

5. Mengembangkan rancangan *evaluation pipeline* yang memungkinkan pelaksanaan eksperimen *multi model* dan *multi persona* secara terotomatisasi.

I.4 Batasan Masalah

Batasan masalah ditetapkan agar ruang lingkup penelitian terkelola dan selaras dengan tujuan penelitian. Penelitian ini tidak bertujuan mengevaluasi seluruh aspek perilaku model bahasa, tetapi fokus pada pengaruh *user persona*. Batasan penelitian ini adalah sebagai berikut.

1. Penelitian hanya menganalisis dua jenis *user persona*, yaitu *user persona* eksplisit dan *user persona* implisit. Penelitian tidak mencakup *role-playing persona* yang memberikan identitas kepada model maupun mekanisme *personalization* berbasis histori pengguna.
2. Pengujian terbatas pada model bahasa berbasis teks yang dapat diakses melalui API. Model multimodal, model yang memerlukan *fine-tuning*, atau model yang memerlukan pelatihan ulang tidak termasuk dalam cakupan penelitian.
3. Evaluasi dibatasi pada tugas berbasis teks, termasuk penalaran numerik, penalaran logis, pertanyaan pengetahuan umum, skenario sosial, dan skenario moral. Tugas vision-language atau *speech* tidak dibahas.
4. Penilaian kualitas keluaran dilakukan melalui evaluasi terotomatisasi dan analisis komparatif. Penilaian berbasis partisipan manusia tidak dilakukan.
5. Penelitian menggunakan *evaluation pipeline* berbasis eksekusi prompt tanpa melakukan modifikasi pada parameter internal model.
6. Analisis bias terbatas pada *human bias* yang muncul sebagai akibat variasi *user persona*, dan tidak mencakup bias makro yang bersumber dari data pelatihan model.

I.5 Metodologi

Metodologi pada tahap penyusunan proposal ini disusun untuk memastikan bahwa proses perumusan masalah, penentuan ruang lingkup penelitian, dan penyusunan kerangka teoretis dilakukan secara sistematis. Metodologi ini tidak mencakup tahapan implementasi eksperimen, yang akan dijabarkan pada Bab III, melainkan berfokus pada kegiatan awal yang diperlukan untuk menghasilkan proposal penelitian yang terarah dan berbasis kajian ilmiah.

I.5.1 Tahap 1: Investigasi Awal dan Pengumpulan Fakta

Tahap awal dilakukan untuk memahami konteks permasalahan dan mengidentifikasi isu ilmiah yang relevan dengan topik penelitian. Langkah yang dilakukan meliputi:

1. Mengidentifikasi fenomena sensitivitas *large language model* terhadap identitas pengguna berdasarkan contoh kasus, laporan empiris, dan temuan penelitian sebelumnya.
2. Meninjau keluaran awal beberapa model bahasa melalui eksplorasi terbatas untuk mengamati indikasi pengaruh *user persona* eksplisit dan *user persona* implisit terhadap penalaran dan gaya respons.
3. Menyimpulkan pola permasalahan yang muncul untuk kemudian dirumuskan sebagai pokok masalah penelitian.

I.5.2 Tahap 2: Pencarian, Pengelompokan, dan Penapisan Literatur

Tahap ini dilakukan untuk memperoleh landasan ilmiah yang kuat dalam menyusun kerangka teoretis dan menentukan arah penelitian. Kegiatan yang dilakukan mencakup:

1. Melakukan pencarian literatur menggunakan mesin pencarian akademik seperti Google Scholar, Semantic Scholar, arXiv, dan ACL Anthology dengan kata kunci antara lain *user persona*, *implicit persona*, *identity-conditioned prompting*, *LLM sensitivity*, *reasoning evaluation*, dan *bias in LLM*.
2. Menyeleksi publikasi yang relevan, termasuk penelitian mengenai pengaruh identitas pengguna terhadap keluaran model bahasa, teori penalaran pada model bahasa, evaluasi berbasis prompt, dan bias implisit.
3. Mengelompokkan literatur ke dalam kategori konseptual, yaitu: (a) konsep dasar *large language model*, (b) teori dan klasifikasi *persona* eksplisit dan implisit, (c) penelitian terdahulu mengenai identitas pengguna dan pengaruhnya terhadap keluaran model, (d) metode evaluasi penalaran dan analisis bias.
4. Menganalisis dan merangkum kontribusi, metodologi, serta keterbatasan setiap publikasi yang terpilih untuk memastikan bahwa kerangka teoretis proposal didasarkan pada referensi yang valid dan mutakhir.
5. Mendokumentasikan seluruh proses penelusuran literatur, termasuk daftar kata kunci, sumber pencarian, dan kriteria penapisan yang digunakan. Dokumentasi tambahan, seperti rekaman proses eksplorasi awal atau catatan observasi, akan dicantumkan pada bagian lampiran.

Tahap-tahap tersebut menghasilkan landasan konseptual dan rumusan permasalahan yang digunakan dalam penyusunan proposal tugas akhir. Hasil kajian literatur secara

rinci akan disajikan pada Bab II Studi Literatur.

BAB II

STUDI LITERATUR

Bab ini membahas konsep dan penelitian terdahulu yang menjadi landasan bagi analisis pengaruh *user persona* terhadap perilaku *large language model*. Pembahasan disusun secara bertahap, dimulai dari uraian mengenai model bahasa modern, mekanisme pemrosesan instruksi, konsep dasar persona, serta temuan empiris mengenai sensitivitas model terhadap identitas pengguna. Selain itu, bab ini meninjau isu bias dan metode evaluasi penalaran yang relevan bagi perancangan penelitian ini.

II.1 Large Language Model

II.1.1 Konsep dan Karakteristik Dasar

Large language model (LLM) merupakan model generatif berbasis arsitektur transformator yang dilatih menggunakan data dalam skala sangat besar. Model ini mempelajari pola bahasa melalui hubungan antartoken, sehingga mampu membangun representasi yang mencakup makna, hubungan semantik, serta isyarat pragmatik yang muncul dalam teks. Dengan skala pelatihan yang luas, LLM dapat digunakan pada berbagai tugas tanpa memerlukan penyesuaian khusus untuk setiap tugas.

Secara konseptual, LLM bekerja dengan memprediksi token berikutnya berdasarkan konteks sebelumnya. Namun, proses prediksi ini tidak sekadar berbasis frekuensi kata, melainkan menggunakan representasi kontekstual yang memungkinkan model memahami instruksi, gaya penulisan, maupun kecenderungan komunikasi. Model seperti GPT, LLaMA, Mistral, dan Gemini mengadopsi pendekatan ini dan menunjukkan kemampuan generalisasi yang kuat terhadap tugas bahasa yang kompleks.

Karakteristik utama LLM antara lain fleksibilitas dalam mengikuti instruksi, kemampuan menyusun penalaran, serta penyesuaian terhadap pola komunikasi pengguna. Kemampuan ini muncul dari kombinasi arsitektur dasar transformator, skala

parameter yang besar, dan keragaman data pelatihan. Karena model tidak dibuat untuk satu domain tertentu, tetapi dilatih pada data lintas konteks, gaya, dan situasi, LLM dapat mengadaptasi perilaku komunikasinya berdasarkan variasi kecil dalam instruksi.

II.1.2 Representasi Bahasa dan Pemahaman Instruksi

LLM memproses teks melalui beberapa tahapan representasi internal. Teks diuraikan menjadi token, kemudian dipetakan ke dalam ruang representasi berdimensi tinggi melalui *embedding*. Representasi awal ini kemudian diperkaya melalui lapisan-lapisan transformator yang memanfaatkan mekanisme perhatian untuk menentukan hubungan antar token dalam konteks yang lebih luas. Hasilnya adalah representasi kontekstual yang mencerminkan interpretasi model terhadap instruksi atau percakapan.

Representasi ini tidak bersifat statis. Makna sebuah token dapat berubah bergantung pada cara pengguna menyampaikan instruksi. Perbedaan gaya penulisan, urutan informasi, atau tingkat formalitas dapat menghasilkan representasi internal yang berbeda, sehingga memunculkan respons yang berbeda pula. Penelitian Zhou et al. (Zhou dkk. 2023) menunjukkan bahwa perubahan kecil dalam framing, seperti perbedaan nada atau cara bertanya, dapat menggeser perhatian model dan mengubah struktur jawaban yang dihasilkan.

Sebagai ilustrasi, perbedaan instruksi berikut sering kali menghasilkan respons yang berbeda meskipun inti pertanyaannya sama:

- "Jelaskan secara singkat apa itu regularisasi."
- "Saya sedang menulis laporan akademik. Bisakah Anda menjelaskan secara formal apa yang dimaksud dengan regularisasi?"

Instruksi kedua biasanya memicu model untuk memberikan penjelasan yang lebih panjang, lebih berhati-hati, dan lebih formal. Perbedaan ini mencerminkan bagaimana representasi instruksi terbentuk berdasarkan konteks linguistik dan pragmatik.

II.1.3 Penalaran dan Dinamika Perilaku Model

Selain pemahaman instruksi, LLM juga menunjukkan kemampuan melakukan penalaran. Model dapat menyelesaikan soal penalaran numerik sederhana, menjawab pertanyaan berbasis pengetahuan umum, hingga memberikan penilaian terhadap skenario sosial atau moral. Namun, kemampuan ini tidak sepenuhnya stabil. Turpin et al. (Turpin dkk. 2023) menemukan bahwa penalaran yang dihasilkan model

dapat berubah hanya karena variasi kecil pada bentuk instruksi, walaupun substansi tugas tetap sama.

Hal ini terjadi karena model tidak melakukan penalaran melalui prosedur logis eksplisit, tetapi melalui dinamika representasi internal yang sensitif terhadap konteks. Sebuah instruksi yang lebih panjang atau lebih formal dapat memicu struktur penalaran yang lebih sistematis, sementara instruksi yang lebih langsung dapat menghasilkan jawaban tanpa uraian langkah-langkah penalaran yang jelas. Perubahan ini memperlihatkan bahwa struktur penalaran yang muncul merupakan fungsi dari konteks interaksi, bukan semata-mata fungsi dari logika masalah yang diberikan.

Ketidakstabilan ini penting untuk dipahami karena berhubungan langsung dengan penelitian mengenai *user persona*. Jika perubahan kecil pada instruksi dapat mengubah penalaran, maka variasi identitas pengguna yang tersirat dalam tulisan juga berpotensi memicu perubahan serupa.

II.1.4 Dimensi Sosial dalam Pemrosesan Bahasa

Model bahasa modern tidak hanya mempelajari struktur dan makna bahasa, tetapi juga pola interaksi sosial yang tercermin dalam data pelatihan. Weidinger et al. (Weidinger dkk. 2021) menunjukkan bahwa LLM dapat menginternalisasi norma sosial, stereotip, serta pola komunikasi yang umum digunakan manusia. Dalam banyak kasus, gaya bahasa tertentu diinterpretasikan sebagai sinyal sosial mengenai siapa pengguna tersebut, misalnya usia, latar profesional, atau tingkat pendidikan.

Ketika instruksi ditulis dengan gaya santai, model sering kali memberikan respons yang lebih ringkas atau lebih langsung. Sebaliknya, ketika instruksi ditulis dengan gaya formal, respons yang dihasilkan cenderung lebih berhati-hati dan mengikuti struktur penjelasan akademis. Perbedaan respons ini bukan sekadar akibat gaya penulisan, tetapi akibat inferensi sosial yang dilakukan model berdasarkan pola komunikasi dalam data pelatihan.

Fenomena ini menunjukkan bahwa pemrosesan bahasa oleh LLM memiliki dimensi sosial yang signifikan. Instruksi diperlakukan bukan hanya sebagai teks, tetapi sebagai bentuk interaksi manusia yang membawa sinyal identitas. Sensitivitas terhadap sinyal ini merupakan salah satu alasan mengapa *user persona* dapat memengaruhi penalaran, struktur respons, maupun kecenderungan bias dalam keluaran model.

II.2 Persona dalam Interaksi Model Bahasa

II.2.1 Definisi dan Ruang Lingkup Persona

Dalam kajian sistem bahasa alami, *persona* merujuk pada serangkaian atribut yang digunakan untuk menggambarkan identitas atau karakteristik pengguna. Atribut tersebut dapat berupa informasi sosial, demografis, profesional, atau gaya komunikasi yang merepresentasikan cara seseorang berinteraksi dalam percakapan. Persona berfungsi sebagai konteks tambahan yang dapat memengaruhi bagaimana sebuah sistem dialog memahami maksud pengguna dan membentuk respons.

Dalam konteks *large language model*, persona tidak hanya dipandang sebagai label identitas, tetapi juga sebagai bagian dari sinyal yang terkandung dalam bahasa. Karena model belajar dari data pelatihan yang mencerminkan cara manusia berkomunikasi, model juga mempelajari keterkaitan antara gaya bahasa dan identitas sosial. Dengan demikian, persona tidak hanya bekerja sebagai informasi eksplisit, tetapi dapat tersirat melalui variasi linguistik seperti pilihan kata, nada, struktur kalimat, atau keformalan tulisan.

Ruang lingkup persona dalam sistem bahasa mencakup berbagai kategori identitas, seperti gender, usia, minat, latar profesional, afiliasi budaya, ataupun preferensi komunikasi. Representasi persona tersebut tidak selalu hadir dalam bentuk pernyataan langsung, tetapi sering kali dinyatakan melalui konteks linguistik yang halus tanpa deklarasi eksplisit mengenai siapa pengguna tersebut.

II.2.2 Persona Eksplisit dan Persona Implisit

Fenomena persona dalam interaksi dengan model bahasa dapat dibagi menjadi dua bentuk utama, yaitu persona eksplisit dan persona implisit. Keduanya memberikan sinyal identitas, tetapi melalui mekanisme dan intensitas yang berbeda.

Persona eksplisit muncul ketika identitas pengguna dinyatakan secara langsung dalam instruksi atau konteks percakapan. Contohnya adalah ketika pengguna menuliskan “Saya adalah mahasiswa teknik informatika” atau “Sebagai seorang dokter, saya ingin memahami...”. Ungkapan seperti ini memberikan sinyal yang jelas kepada model mengenai latar pengguna, sehingga model dapat menyesuaikan struktur respons agar lebih sesuai dengan karakteristik tersebut. Gupta et al. (Gupta dkk. 2024) menunjukkan bahwa penugasan persona eksplisit semacam ini dapat mengubah hasil penalaran model, meskipun tugas yang diberikan tidak berkaitan dengan identitas sosial pengguna. Perubahan respons tidak hanya menyangkut gaya bahasa,

tetapi juga dapat memengaruhi kesimpulan logis yang diberikan model.

Sebaliknya, persona implisit muncul ketika identitas pengguna tidak dinyatakan secara langsung, tetapi disimpulkan oleh model berdasarkan isyarat linguistik. Penelitian Tseng et al. (Tseng dkk. 2024) menunjukkan bahwa model memiliki kecenderungan melakukan inferensi identitas pengguna dari gaya penulisan, struktur kalimat, pilihan kata, atau tingkat formalitas. Fenomena ini dapat terjadi meskipun pengguna tidak bermaksud menyampaikan identitas tertentu. Sebagai contoh, gaya penulisan formal dengan istilah akademis sering diasosiasikan dengan latar pendidikan tertentu, sedangkan gaya penulisan santai dapat diasosiasikan dengan kategori usia atau tingkat kedekatan sosial.

Inferensi identitas tersebut bukan hasil dari aturan yang ditetapkan secara eksplisit dalam model, tetapi merupakan konsekuensi dari pola komunikasi manusia yang terserap selama proses pelatihan. Model mempelajari bahwa gaya bahasa tertentu sering muncul bersama atribut sosial tertentu, sehingga ketika gaya tersebut muncul dalam instruksi, model cenderung mengaktifkan pola respons yang sesuai dengan kategori identitas yang diasosiasikan. Fenomena ini menjadi dasar penting bagi studi mengenai pengaruh persona implisit terhadap perilaku dan penalaran model.

II.2.3 Peran Persona dalam Interaksi dengan LLM

Persona, baik eksplisit maupun implisit, berperan sebagai sinyal kontekstual yang memengaruhi interpretasi dan respons model bahasa. Ketika identitas pengguna muncul dalam bentuk atribut sosial atau gaya komunikasi tertentu, model akan memperlakukannya sebagai bagian dari konteks yang relevan. Konteks ini kemudian membentuk representasi internal yang memengaruhi bagaimana model memahami pertanyaan, menafsirkan maksud, dan menyusun jawaban.

Peran persona dalam interaksi ini dapat dilihat dari dua dimensi utama. Pertama, persona dapat memengaruhi aspek linguistik respons, seperti pilihan kata, tingkat formalitas, pola argumentasi, atau struktur penjelasan. Model cenderung menyesuaikan respons agar selaras dengan gaya komunikasi yang diasosiasikan dengan persona tertentu. Kedua, persona dapat memengaruhi penalaran model melalui apa yang disebut sebagai *reasoning shift*, yaitu perubahan struktur penalaran yang terjadi akibat variasi identitas pengguna meskipun subtansi tugas tetap sama.

Sebagai ilustrasi, suatu pertanyaan logika sederhana yang diajukan oleh pengguna dengan persona profesional tertentu dapat memicu model untuk memberikan res-

pons yang lebih sistematis atau lebih berhati-hati. Sebaliknya, pertanyaan yang diajukan dengan gaya informal dapat menghasilkan respons yang lebih ringkas dengan struktur penalaran minimal. Perubahan ini menunjukkan bahwa persona berfungsi sebagai variabel kondisi yang membentuk dinamika interaksi antara pengguna dan model.

II.3 Pengaruh Persona terhadap Perilaku LLM

Pembahasan mengenai persona tidak berhenti pada bagaimana identitas pengguna direpresentasikan dalam instruksi, tetapi juga mencakup bagaimana identitas tersebut memengaruhi perilaku model bahasa ketika menghasilkan respons. Berbagai penelitian menunjukkan bahwa persona berperan sebagai konteks tambahan yang secara halus membentuk cara model memahami pertanyaan, menimbang informasi, dan menyusun jawaban. Dengan demikian, persona tidak sekadar menjadi atribut linguistik, tetapi menjadi bagian dari dinamika interaksi yang memengaruhi proses penalaran dan karakter keluaran model.

II.3.1 Pengaruh Persona terhadap Penalaran Model

Penalaran merupakan salah satu kemampuan utama yang ditonjolkan oleh model bahasa modern. Namun, sejumlah studi menemukan bahwa penalaran tersebut tidak selalu stabil dan dapat berubah bergantung pada konteks identitas pengguna. Gupta et al. (Gupta dkk. 2024) menunjukkan bahwa ketika sebuah persona eksplisit disisipkan ke dalam instruksi, model dapat menghasilkan struktur penalaran yang berbeda meskipun tugas yang diberikan tetap sama. Perubahan tersebut terlihat pada pemilihan langkah-langkah argumentatif, urutan penjelasan, atau tingkat kehati-hatian dalam menarik kesimpulan.

Dalam konteks persona implisit, perubahan penalaran muncul melalui mekanisme yang lebih halus. Gaya penulisan pengguna, seperti tingkat formalitas, panjang kalimat, atau pilihan kosakata, dapat diinterpretasikan sebagai sinyal identitas yang memengaruhi cara model membangun penalaran. Misalnya, instruksi yang disampaikan dengan gaya akademis sering kali mendorong model untuk memberikan penjelasan yang lebih sistematis dan rinci. Sebaliknya, instruksi yang ditulis dengan gaya santai dapat menghasilkan penalaran yang lebih ringkas atau langsung.

Temuan-temuan ini sejalan dengan penelitian mengenai ketidakstabilan penalaran yang dilakukan oleh Turpin et al. (Turpin dkk. 2023). Dalam studi tersebut, perubahan kecil pada struktur instruksi terbukti memengaruhi urutan *chain-of-thought*

yang dihasilkan model. Karena persona bekerja sebagai bagian dari konteks instruksi, variasi identitas pengguna berpotensi menimbulkan pergeseran pola berpikir yang muncul dalam respons model.

Pengaruh persona terhadap penalaran tampak pada berbagai kategori tugas, mulai dari penalaran numerik hingga pertimbangan moral. Pada tugas numerik, persona tertentu dapat mendorong model untuk memberikan uraian langkah yang lebih panjang atau lebih hati-hati. Pada tugas logika, persona dapat memengaruhi cara model menyusun argumen. Sementara itu, pada tugas sosial atau moral, persona dapat mengarahkan model untuk menekankan nilai-nilai tertentu atau memilih perspektif yang lebih dekat dengan identitas pengguna yang diasumsikan.

II.3.2 Pengaruh Persona terhadap Gaya dan Struktur Respons

Selain penalaran, persona juga memengaruhi aspek gaya dan struktur respons. Model bahasa modern tidak hanya menghasilkan jawaban berdasarkan isi pertanyaan, tetapi juga menyesuaikan cara penyampaiannya agar selaras dengan identitas pengguna yang terdeteksi. Temuan Tseng et al. (Tseng dkk. 2024) menunjukkan bahwa model dapat meniru gaya bahasa yang diasosiasikan dengan persona tertentu, bahkan ketika identitas tersebut tidak dinyatakan secara eksplisit.

Perubahan yang muncul dapat berupa pemilihan kosakata, panjang penjelasan, tingkat formalitas, atau nada yang digunakan dalam respons. Apabila model mengaitkan pengguna dengan latar profesional tertentu, respons yang dihasilkan sering kali lebih teknis atau lebih terstruktur. Sebaliknya, apabila gaya penulisan pengguna menunjukkan kedekatan sosial atau informalitas, respons yang muncul cenderung lebih ringkas atau lebih langsung.

Dalam beberapa kasus, persona tertentu juga dapat memicu model untuk bersikap lebih berhati-hati, terutama pada topik-topik yang sensitif secara sosial. Fenomena ini muncul karena model mempelajari pola komunikasi manusia dalam data pelatihan dan mengaitkan gaya bahasa dengan norma sosial yang berlaku pada kelompok tertentu. Dengan demikian, perbedaan gaya respons bukan sekadar variasi permukaan, tetapi merupakan hasil dari proses interpretasi sosial yang dilakukan model.

II.3.3 Faktor yang Memperkuat Efek Persona

Variasi respons akibat persona diperkuat oleh sejumlah faktor yang berkaitan dengan konteks interaksi. Salah satu faktor tersebut adalah framing instruksi. Ketika persona disampaikan secara konsisten, baik melalui deskripsi eksplisit maupun ga-

ya penulisan yang stabil, representasi identitas pengguna menjadi lebih kuat dalam interpretasi model. Hal ini membuat model lebih cenderung mempertahankan pola respons tertentu sepanjang percakapan.

Selain itu, jenis tugas yang diberikan turut memengaruhi seberapa besar dampak persona terhadap respons model. Tugas yang bersifat terbuka, seperti pertanyaan moral atau skenario sosial, memberikan ruang interpretasi yang lebih luas sehingga sinyal identitas lebih mudah memengaruhi pola jawaban. Sebaliknya, tugas-tugas yang memiliki jawaban pasti atau struktur penyelesaian yang ketat cenderung menunjukkan pengaruh persona yang lebih kecil.

Skala model dan metode penyelarasan instruksi juga memainkan peran penting. Model yang dilatih dengan data percakapan dalam jumlah besar cenderung lebih sensitif terhadap variasi gaya linguistik. Sementara itu, model dengan kapasitas lebih kecil dapat menunjukkan respons yang kurang konsisten karena representasi sosial yang terbatas.

Secara keseluruhan, efek persona merupakan hasil interaksi antara konteks linguistik, representasi sosial, dan mekanisme penyelarasan model. Faktor-faktor ini bekerja bersamaan dan membentuk variasi respons yang menggambarkan bagaimana model bahasa menafsirkan identitas pengguna dalam proses menghasilkan jawaban.

II.4 Bias dalam Respons LLM

Pembahasan mengenai bias dalam *large language model* berangkat dari kenyataan bahwa model bahasa belajar dari pola-pola yang muncul dalam data pelatihan. Data tersebut bukan hanya berisi informasi faktual, tetapi juga memuat kecenderungan sosial yang terbentuk secara historis. Ketika model mempelajari pola bahasa dari data tersebut, model tidak hanya menyerap struktur linguistik, tetapi juga asumsi-asumsi sosial yang secara tidak sengaja dapat tercermin dalam respons yang dihasilkan. Dalam konteks penelitian ini, bias menjadi penting karena persona—baik eksplisit maupun implisit—dapat memperkuat atau menggeser pola bias yang dimiliki model.

II.4.1 Bentuk-bentuk Bias pada Model Bahasa

Bias dalam model bahasa dapat muncul dalam beragam bentuk. Salah satu bentuk yang sering menjadi perhatian adalah bias representasional, yaitu kecenderungan model menggambarkan suatu kelompok sosial secara tidak seimbang. Weidinger et

al. (Weidinger dkk. 2021) menunjukkan bahwa stereotip yang sering muncul dalam teks internet dapat terinternalisasi dalam model. Misalnya, model dapat mengaitkan profesi tertentu dengan gender tertentu atau menempatkan kelompok sosial tertentu dalam peran tertentu, meskipun konteks yang diberikan sebenarnya netral.

Selain bias representasional, terdapat bias inferensial, yaitu kecenderungan model mengambil kesimpulan berdasarkan isyarat yang tidak relevan. Bentuk bias ini biasanya muncul ketika model meniru pola asosiasi dari data tanpa memahami konteks sebenarnya. Sebagai contoh, ketika diminta mendeskripsikan seseorang dalam skenario imajinatif, model dapat mengisi detail yang tidak disebutkan hanya karena mengikuti pola umum yang sering ditemui dalam data pelatihan.

Bias tersebut tidak hanya muncul pada isi jawaban, tetapi juga dalam cara model menyusun uraian penjelasan. Pada topik-topik moral atau sosial, bias dapat terlihat dari pilihan nilai atau asumsi yang digunakan dalam penalaran. Hal ini memperlihatkan bahwa bias dalam model bahasa bersifat berlapis: ia dapat memengaruhi kosakata, struktur kalimat, hingga cara model melakukan evaluasi terhadap suatu situasi.

II.4.2 Konsekuensi Bias terhadap Keluaran Model

Bias yang muncul dalam model bahasa membawa sejumlah konsekuensi terhadap keluaran yang diberikan kepada pengguna. Salah satu konsekuensi yang paling sering dibahas adalah risiko misinformasi. Ketika model memberikan jawaban yang terdengar meyakinkan tetapi sebenarnya bias atau tidak akurat, pengguna yang tidak memiliki pengetahuan memadai dapat menerima informasi tersebut sebagai kebenaran.

Konsekuensi lainnya berkaitan dengan ketidakmerataan kualitas respons. Jika model menyesuaikan gaya penjelasan berdasarkan persona tertentu, kelompok pengguna yang berbeda dapat menerima penjelasan dengan tingkat kedalaman atau kehatihatian yang tidak sama. Walaupun model tidak memiliki niat atau tujuan tertentu, perbedaan kualitas informasi ini dapat mempengaruhi proses pemahaman pengguna terhadap suatu topik.

Di sisi lain, bias juga dapat memperkuat stereotip sosial. Ketika model berulang kali memberikan deskripsi atau penilaian yang sejalan dengan stereotip tertentu, model secara tidak langsung ikut berpartisipasi dalam memperkuat persepsi sosial yang tidak akurat. Penguatan stereotip ini dapat terjadi secara halus, misalnya

melalui pilihan kosakata yang cenderung bernuansa tertentu atau struktur argumen yang mengarah pada penilaian yang bias.

II.4.3 Kaitannya dengan Variasi Persona

Persona, sebagai sinyal identitas pengguna, dapat memperkuat atau menggeser munculnya bias dalam respons model. Pada persona eksplisit, bias dapat timbul ketika model mengaitkan identitas pengguna dengan pola stereotip dalam data pelatihan. Misalnya, pernyataan seperti "Saya seorang guru" atau "Saya berasal dari profesi X" dapat memicu model memberikan respons yang mengikuti pola tertentu yang sering dikaitkan dengan profesi tersebut.

Pada persona implisit, bias muncul dengan cara yang lebih halus. Karena model sangat peka terhadap gaya penulisan, pilihan kata atau tingkat formalitas dapat dianggap sebagai indikator identitas sosial pengguna. Jika model mengaitkan gaya komunikasi tertentu dengan kelompok sosial tertentu, respons yang dihasilkan dapat mencerminkan bias yang dimiliki model terhadap kelompok tersebut.

Fenomena ini menjadi lebih terlihat pada tugas-tugas yang bersifat terbuka, seperti pertanyaan moral, skenario etika, atau pertimbangan sosial. Pada jenis tugas tersebut, respons model sangat dipengaruhi oleh konteks dan cara model melakukan inferensi sosial. Ketika persona menjadi bagian dari konteks, respons yang dihasilkan dapat menunjukkan pergeseran nilai, perhatian, atau prioritas tertentu. Kondisi inilah yang membuat analisis persona dalam penelitian ini menjadi penting: persona tidak hanya memengaruhi gaya bahasa atau cara penalaran, tetapi juga membuka ruang bagi bias untuk muncul atau berubah.

II.5 Evaluasi Penalaran dan Benchmark

Evaluasi terhadap *large language model* tidak hanya dilakukan dengan melihat kemampuan model menghasilkan teks, tetapi juga melalui serangkaian tugas terstruktur yang dirancang untuk mengukur kemampuan penalaran, pemahaman konteks, serta kemampuan model menyelesaikan masalah secara konsisten. Benchmark menjadi alat penting dalam penelitian karena memberikan gambaran yang lebih objektif mengenai bagaimana model berperilaku di berbagai situasi dan tingkat kesulitan. Dalam konteks penelitian ini, benchmark yang digunakan tidak hanya berfungsi untuk menilai performa penalaran, tetapi juga untuk melihat bagaimana persona dapat memengaruhi keluaran model pada berbagai jenis tugas.

II.5.1 Benchmark Penalaran dan Pengetahuan

Sejumlah benchmark telah dikembangkan untuk mengukur kemampuan penalaran model bahasa. Salah satu yang paling dikenal adalah GSM8K, sebuah kumpulan soal matematika tingkat sekolah dasar yang dirancang untuk menguji penalaran numerik dan kemampuan model menyusun langkah-langkah penyelesaian secara terstruktur. Meskipun soalnya sederhana bagi manusia, benchmark ini cukup menantang bagi model bahasa karena mengharuskan model memahami konteks, menerapkan logika dasar, dan menjaga konsistensi antara uraian langkah dan jawaban akhir.

Selain GSM8K, benchmark lain seperti MMLU digunakan untuk menguji kemampuan model pada pertanyaan lintas domain, mulai dari sains hingga ilmu sosial. MMLU menekankan kapasitas model dalam memahami pengetahuan faktual dan menerapkannya dalam konteks yang tepat. Benchmark ini memberikan gambaran mengenai seberapa baik model dapat menjawab pertanyaan yang membutuhkan pemahaman konsep dan penalaran tingkat menengah.

Benchmark semacam ini penting karena menampilkan kemampuan dasar model tanpa dipengaruhi oleh gaya interaksi yang terlalu terbuka. Dengan kata lain, benchmark berbasis pengetahuan atau logika dasar memberikan titik awal yang netral sebelum mempertimbangkan bagaimana persona dapat menggeser atau memengaruhi jawaban model.

II.5.2 Benchmark Sosial dan Moral

Di samping penalaran numerik dan faktual, kemampuan model untuk memahami situasi sosial dan moral juga menjadi perhatian dalam penelitian. Benchmark seperti SocialIQA digunakan untuk mengukur kemampuan model memahami skenario sosial sederhana, misalnya bagaimana seseorang mungkin merespons suatu tindakan atau apa motivasi yang mungkin dimiliki dalam konteks tertentu. Benchmark ini menekankan bagaimana model menginternalisasi pola interaksi antarindividu berdasarkan data pelatihan.

Selain SocialIQA, terdapat pula tugas-tugas moral yang dirancang untuk melihat bagaimana model memberikan penilaian terhadap situasi etis. Tugas semacam ini tidak memiliki jawaban pasti, sehingga respons model sangat dipengaruhi oleh nilai, norma, atau pola argumentasi yang diserap selama pelatihan. Dalam konteks penelitian persona, tugas moral menjadi menarik karena persona dapat menggeser sudut pandang moral yang diambil model, misalnya apakah model menjadi lebih

berhati-hati, lebih permissive, atau lebih normatif.

Benchmark sosial dan moral ini penting untuk menganalisis bagaimana persona bekerja pada situasi yang tidak memiliki jawaban tunggal dan mengharuskan model melakukan interpretasi berdasarkan konteks sosial.

II.5.3 Tantangan Evaluasi Berbasis Persona

Meskipun benchmark merupakan alat penting dalam evaluasi model, penggunaan benchmark dalam penelitian persona memiliki tantangan tersendiri. Salah satu tantangan utama adalah konsistensi. Karena persona dapat memengaruhi gaya penalaran dan respons model, evaluasi harus dilakukan dengan cara yang memastikan bahwa perubahan yang muncul benar-benar disebabkan oleh persona, bukan oleh variasi lain dalam instruksi atau struktur prompt.

Tantangan berikutnya adalah sensitivitas model terhadap framing. Perubahan kecil pada instruksi, bahkan ketika persona tidak berubah, dapat menghasilkan respons yang berbeda. Hal ini membuat evaluasi berbasis persona memerlukan desain eksperimen yang hati-hati agar pengaruh persona dapat dipisahkan dari pengaruh variasi linguistik.

Selain itu, model bahasa cenderung mengalami *drift* atau perubahan perilaku kecil antarevaluasi, terutama jika evaluasi tidak terotomatisasi dengan baik. Hal ini dapat memengaruhi replikasi hasil dan interpretasi terhadap pengaruh persona. Penggunaan *pipeline* evaluasi yang terstandardisasi dapat membantu mengurangi variasi ini dengan memastikan bahwa setiap model menerima struktur instruksi yang konsisten.

Secara keseluruhan, benchmark memberikan fondasi penting untuk memahami perilaku model dalam berbagai skenario. Namun, dalam konteks penelitian persona, benchmark tidak hanya berfungsi sebagai alat ukur kemampuan, tetapi juga sebagai sarana untuk melihat bagaimana identitas pengguna dapat menggeser proses penalaran, struktur respons, dan kualitas informasi yang diberikan model.

II.6 Penelitian Terdahulu dan Kesenjangan Penelitian

Pembahasan mengenai persona dan perilaku model bahasa telah menjadi bagian dari diskusi yang semakin luas dalam penelitian model berbasis transformator. Sejumlah studi sebelumnya memberikan gambaran mengenai bagaimana identitas pengguna, baik yang dinyatakan secara eksplisit maupun tersirat melalui gaya penulisan, dapat

memengaruhi respons model. Meskipun demikian, penelitian-penelitian tersebut umumnya memiliki cakupan yang terbatas pada satu jenis persona, satu kategori tugas, atau satu model tertentu. Bagian ini merangkum temuan utama dari penelitian terdahulu serta mengidentifikasi sejumlah kesenjangan yang melatarbelakangi penyusunan penelitian ini.

II.6.1 Ringkasan Literatur Terkait

Gupta et al. (Gupta dkk. 2024) menunjukkan bahwa persona eksplisit yang diberikan dalam instruksi dapat mengubah struktur penalaran model, bahkan ketika tugas yang diberikan tidak berkaitan dengan identitas sosial tersebut. Temuan ini membuka diskusi bahwa model tidak hanya memproses isi instruksi, tetapi juga memaknai identitas sebagai konteks tambahan yang membentuk langkah-langkah penalaran.

Di sisi lain, Tseng et al. (Tseng dkk. 2024) menyoroti fenomena persona implisit yang muncul dari gaya penulisan pengguna. Model dapat menafsirkan pilihan kata, tingkat formalitas, atau cara menyampaikan pertanyaan sebagai sinyal identitas, sehingga menghasilkan respons yang selaras dengan kategori sosial yang diasosiasikan dengan isyarat tersebut. Studi ini menunjukkan bahwa persona tidak harus dinyatakan secara eksplisit untuk memengaruhi respons model.

Penelitian lain menyoroti aspek ketidakstabilan penalaran model. Turpin et al. (Turpin dkk. 2023) menunjukkan bahwa perubahan kecil pada struktur instruksi dapat mengubah langkah *chain-of-thought* yang dihasilkan model. Hal ini menunjukkan bahwa proses penalaran model sangat bergantung pada konteks linguistik, termasuk gaya atau nada instruksi yang pada akhirnya berhubungan erat dengan persona.

Dalam konteks bias, Weidinger et al. (Weidinger dkk. 2021) menunjukkan bahwa model bahasa dapat memperkuat atau meniru pola stereotip yang ada dalam data pelatihan. Temuan ini relevan ketika dikaitkan dengan persona karena identitas pengguna dapat memperkuat pola bias tertentu, terutama pada tugas sosial dan moral yang melibatkan interpretasi nilai atau pengambilan posisi tertentu.

Penelitian mengenai personalisasi model, seperti yang dibahas dalam Naous et al. (Naous, Roziere, dkk. 2025), lebih menekankan bagaimana variasi preferensi pengguna dapat memengaruhi gaya atau struktur jawaban. Meskipun fokusnya berbeda, studi ini memberikan gambaran bahwa model bahasa memberikan respons yang bervariasi bergantung pada konteks identitas atau preferensi pengguna.

II.6.2 Keterbatasan Penelitian Sebelumnya

Meskipun penelitian-penelitian tersebut memberikan kontribusi penting dalam memahami hubungan antara persona dan perilaku model bahasa, sebagian besar studi masih memiliki sejumlah keterbatasan. Pertama, banyak penelitian hanya menggunakan satu model sehingga temuan yang diperoleh belum menggambarkan variasi perilaku antarmodel. Padahal, model yang berbeda dapat menunjukkan tingkat sensitivitas yang berbeda terhadap persona.

Kedua, cakupan persona yang diteliti cenderung terbatas, sering kali hanya mencakup beberapa persona eksplisit atau sejumlah contoh persona implisit yang relatif kecil. Kondisi ini membuat temuan penelitian sebelumnya belum cukup untuk menggambarkan bagaimana variasi persona yang lebih luas memengaruhi perilaku model.

Ketiga, sebagian besar penelitian hanya menguji satu atau dua jenis tugas. Padahal, persona dapat memengaruhi model secara berbeda pada penalaran numerik, penalaran logis, skenario sosial, maupun pertanyaan moral. Keterbatasan cakupan tugas ini membuat analisis sebelumnya belum mencerminkan penuh kompleksitas pengaruh persona.

Keempat, sebagian penelitian belum menyediakan kerangka evaluasi yang terotomatisasi dan konsisten. Tanpa mekanisme evaluasi yang terstruktur, sulit untuk memastikan bahwa perubahan respons benar-benar disebabkan oleh persona dan bukan oleh variasi lain seperti perbedaan prompt atau kejadian *drift* antarpernyataan.

II.6.3 Posisi dan Kontribusi Penelitian Ini

Penelitian ini disusun dengan mempertimbangkan keterbatasan-keterbatasan tersebut. Berbeda dengan penelitian sebelumnya, penelitian ini menggunakan pendekatan *multi model* dan *multi persona* untuk melihat bagaimana variasi identitas pengguna memengaruhi penalaran, gaya respons, dan kecenderungan bias. Dengan mengombinasikan beberapa kategori tugas—mulai dari penalaran numerik hingga skenario moral—penelitian ini bertujuan memberikan gambaran yang lebih utuh mengenai perilaku model bahasa ketika berinteraksi dengan berbagai persona pengguna.

Penelitian ini juga memanfaatkan *evaluation pipeline* yang terotomatisasi untuk memastikan konsistensi struktur instruksi dan mengurangi pengaruh variasi yang tidak diinginkan. Pendekatan ini diharapkan dapat memberikan hasil yang lebih stabil dan dapat direplikasi, sehingga memperkuat kontribusi penelitian dalam memahami

sensitivitas model bahasa terhadap persona.

Secara keseluruhan, penelitian-penelitian tersebut menunjukkan perlunya kajian yang lebih luas dan terstruktur mengenai bagaimana persona memengaruhi perilaku model bahasa, terutama ketika melibatkan lebih dari satu model dan lebih dari satu kategori tugas.

BAB III

ANALISIS MASALAH

III.1 Analisis Kondisi Saat Ini

Perkembangan *large language model* (LLM) dalam beberapa tahun terakhir mendorong pemanfaatan model bahasa dalam berbagai konteks, mulai dari penjawab pertanyaan, agen percakapan, hingga sistem pendukung pengambilan keputusan (Bommasani, Hudson, Adeli, dkk. 2021). Seiring dengan meluasnya penggunaan tersebut, muncul kebutuhan untuk memahami bagaimana model bereaksi terhadap variasi identitas dan karakteristik pengguna, bukan hanya terhadap instruksi tugas. Hal ini berkaitan dengan cara model memproses konteks interaksi yang memuat informasi tentang siapa yang berinteraksi dengan model, dalam kapasitas apa, dan dengan gaya komunikasi seperti apa.

Penelitian mengenai persona pada LLM sejauh ini banyak berfokus pada pemberian identitas kepada model sebagai agen percakapan. Tseng et al. mengkaji berbagai pendekatan *role-playing* dan *personalization* yang umumnya memposisikan persona pada sisi model, misalnya melalui instruksi sistem yang mendeskripsikan karakter, gaya bicara, atau peran yang harus diambil oleh model (Tseng dkk. 2024). Pada pengaturan ini, model diminta untuk bertindak sebagai tenaga profesional, tokoh tertentu, atau asisten dengan gaya komunikasi spesifik, dan evaluasi dilakukan dengan melihat konsistensi gaya respons maupun kesesuaian perilaku dengan persona yang diberikan.

Di luar *role-playing* tersebut, sejumlah studi menunjukkan bahwa penyisipan persona eksplisit dapat memengaruhi penalaran model bahkan pada tugas yang dirancang sebagai soal penalaran abstrak dan tidak secara eksplisit memuat dimensi sosial. Gupta et al. menunjukkan bahwa identitas yang dilekatkan pada konteks dapat menggeser cara model melakukan penalaran dan memilih jawaban, termasuk pada soal yang dirancang untuk menguji penalaran formal (Gupta dkk. 2024). Temuan

ini mengindikasikan bahwa persona tidak hanya memengaruhi gaya bahasa, tetapi juga struktur langkah penalaran yang dihasilkan model.

Pada saat yang sama, struktur penalaran LLM terbukti sensitif terhadap variasi kecil pada instruksi. Turpin et al. memperlihatkan bahwa perubahan ringan dalam formulasi *prompt* dapat menghasilkan rantai penalaran yang berbeda meskipun pertanyaannya sama (Turpin dkk. 2023). Studi lain mengenai sensitivitas model terhadap framing dan gaya penulisan menunjukkan bahwa cara sebuah instruksi disusun dapat memengaruhi isi dan gaya jawaban (Zhou dkk. 2023). Kondisi ini membuat analisis persona menjadi lebih kompleks, karena persona, framing, dan gaya bahasa sering kali hadir secara bersamaan di dalam konteks interaksi, sehingga sulit memisahkan pengaruh masing-masing faktor.

Isu bias menambah lapisan kompleksitas dalam memahami perilaku model. Weidinger et al. menunjukkan bahwa LLM dapat mereproduksi dan memperkuat pola bias sosial yang tercermin dalam data pelatihan (Weidinger dkk. 2021). Ketika identitas sosial tertentu, misalnya terkait gender, profesi, atau latar budaya, dimasukkan ke dalam konteks, respons model berpotensi mencerminkan bias representasional maupun inferensial yang sudah tertanam di dalam parameter model. Dalam konteks persona, hal ini berarti bahwa perbedaan respons akibat variasi identitas pengguna tidak selalu mencerminkan perubahan kemampuan penalaran, tetapi juga dapat berkaitan dengan bias yang telah terinternalisasi.

Sebagian besar studi persona yang ada menempatkan persona pada sisi model, bukan pada sisi pengguna. Instruksi yang mengubah peran model sebagai agen percakapan berbeda dengan skenario di mana konteks interaksi menyatakan bahwa pengguna memiliki identitas atau latar belakang tertentu. Riset mengenai pemodelan pengguna mulai berkembang, misalnya melalui pendekatan *user language model* yang mempelajari distribusi bahasa berdasarkan karakteristik pengguna (Naous, Roziere, dkk. 2025), tetapi penelitian yang secara sistematis mengkaji dampak *user persona* eksplisit maupun implisit terhadap penalaran dan kualitas jawaban pada berbagai tugas masih relatif terbatas.

Dari sisi infrastruktur evaluasi, banyak studi sebelumnya masih mengandalkan eksekusi manual atau setengah otomatis ketika menjalankan eksperimen yang melibatkan variasi pengguna. Naous et al. menyoroti pentingnya pendekatan yang lebih terstruktur ketika mengevaluasi model dalam konteks variasi pengguna, termasuk pengelolaan konfigurasi, pencatatan hasil, serta konsistensi skenario pengujian (Naous, Roziere, dkk. 2025). Tanpa kerangka evaluasi yang terdokumentasi dengan

jelas, eksperimen yang melibatkan banyak model, banyak persona, dan berbagai jenis tugas menjadi sulit direplikasi dan rawan ketidakkonsistenan.

Berdasarkan kondisi tersebut, masalah-masalah utama yang mendasari perumusan penelitian ini dapat diringkas pada Tabel III.1.

Tabel III.1 Daftar masalah penelitian terkait *user persona* pada LLM

Kode	Uraian masalah	Dampak terhadap penelitian
M-01	Persona pada LLM umumnya diterapkan pada sisi model, bukan pada sisi pengguna.	Belum ada pemahaman yang sistematis mengenai bagaimana *user persona* eksplisit maupun implisit memengaruhi penalaran dan kualitas jawaban pada berbagai tugas.
M-02	Efek persona sulit dipisahkan dari efek framing dan gaya penulisan *prompt*.	Perubahan performa atau pola penalaran dapat berasal dari variasi formulasi instruksi, bukan semata akibat perubahan *user persona*, sehingga interpretasi hasil menjadi tidak pasti.
M-03	LLM membawa bias sosial yang terinternalisasi dari data pelatihan.	Ketika identitas pengguna memuat atribut sosial tertentu, respons model berpotensi mencerminkan bias representasional maupun inferensial, sehingga perbedaan jawaban bisa berkaitan dengan bias yang sudah ada di model.
M-04	Cakupan model dan tugas pada studi terdahulu masih terbatas.	Analisis sensitivitas terhadap persona sering kali hanya mencakup sedikit model atau jenis tugas, sehingga belum memberikan gambaran yang cukup luas mengenai variasi perilaku LLM di berbagai konteks.

Masalah M-01 berkaitan dengan dominasi pendekatan yang menempatkan persona pada sisi model. Tseng et al. membahas bagaimana persona digunakan untuk mengubah peran dan gaya respons model melalui instruksi sistem atau deskripsi karakter (Tseng dkk. 2024). Pendekatan ini berbeda dengan skenario di mana identitas dan karakteristik pengguna dinyatakan secara eksplisit atau implisit pada konteks interaksi. Akibatnya, pengaruh *user persona* terhadap penalaran dan kualitas jawaban belum banyak dikaji secara sistematis.

Masalah M-02 muncul karena struktur penalaran LLM sangat sensitif terhadap variasi kecil dalam formulasi instruksi. Turpin et al. menunjukkan bahwa perubah-

an ringan pada susunan *prompt* dapat menghasilkan rantai penalaran yang berbeda meskipun pertanyaannya sama (Turpin dkk. 2023). Zhou et al. juga menunjukkan bahwa framing dan gaya penulisan instruksi dapat memengaruhi isi dan gaya jawaban (Zhou dkk. 2023). Dalam konteks ini, efek *user persona* berpotensi tercampur dengan efek framing, sehingga diperlukan desain eksperimen yang mampu membedakan keduanya.

Masalah M-03 berhubungan dengan bias sosial yang sudah tertanam di dalam model. Weidinger et al. menunjukkan bahwa LLM dapat mereproduksi dan memperkuat pola bias dari data pelatihan (Weidinger dkk. 2021). Ketika *user persona* memuat atribut sosial seperti gender, profesi, atau latar budaya, respons model terhadap persona tersebut dapat dipengaruhi oleh bias yang telah ada sebelumnya. Hal ini menyulitkan interpretasi hasil, karena perbedaan jawaban bisa berasal dari kombinasi antara penyesuaian terhadap persona dan bias yang sudah terinternalisasi di dalam model.

Masalah M-04 menyoroti keterbatasan cakupan model dan tugas pada studi-studi terdahulu. Banyak penelitian persona hanya menguji sedikit model atau fokus pada satu jenis tugas, sehingga belum memberikan gambaran yang cukup luas mengenai bagaimana variasi *user persona* memengaruhi perilaku model pada spektrum tugas penalaran dan percakapan yang lebih beragam (Gupta dkk. 2024; Tseng dkk. 2024). Keterbatasan ini membuka peluang untuk merancang eksperimen yang melibatkan kombinasi multi model dan multi persona pada beberapa kategori tugas yang terpilih.

III.2 Analisis Kebutuhan

Bagian ini menjabarkan kebutuhan penelitian yang diturunkan dari masalah M-01 sampai M-04 pada analisis kondisi saat ini. Kebutuhan tersebut mencakup kebutuhan konseptual dan teknis yang harus dipenuhi agar eksperimen mengenai pengaruh *user persona* eksplisit dan implisit terhadap penalaran, kualitas jawaban, dan kecenderungan *human bias* pada beberapa *large language model* dapat dilaksanakan secara terstruktur.

III.2.1 Identifikasi Masalah Pengguna

Dalam konteks tugas akhir ini, pengguna yang dimaksud adalah peneliti yang ingin mengevaluasi perilaku model bahasa di bawah variasi *user persona*. Berdasarkan analisis pada Bagian Analisis Kondisi Saat Ini, beberapa permasalahan yang diha-

dapi pengguna dapat diidentifikasi sebagai berikut.

1. Definisi dan pengorganisasian *user persona* eksplisit dan *user persona* implisit belum terdokumentasi secara terstruktur. Sebagian besar contoh yang tersedia berfokus pada persona di sisi model, sehingga perumusan persona di sisi pengguna harus disusun sendiri oleh peneliti.
2. Perbedaan keluaran model berpotensi dipengaruhi oleh variasi formulasi instruksi dan framing *prompt*, sehingga tidak selalu jelas apakah perubahan respons model disebabkan oleh variasi *user persona* atau oleh perubahan cara pertanyaan disampaikan.
3. Eksperimen yang melibatkan beberapa model dan beberapa jenis tugas menuntut adanya cara yang terkelola untuk menjalankan skenario yang sama dan mencatat hasilnya secara konsisten, agar dapat dilakukan analisis perbandingan yang sistematis.

Permasalahan-permasalahan tersebut menjadi dasar penyusunan kebutuhan fungsional dan kebutuhan nonfungsional pada penelitian ini.

III.2.2 Kebutuhan Fungsional

Kebutuhan fungsional menggambarkan kemampuan utama yang harus didukung oleh rancangan eksperimen agar permasalahan pada subbagian sebelumnya dapat ditangani. Ringkasan kebutuhan fungsional ditunjukkan pada Tabel III.2.

Tabel III.2 Kebutuhan fungsional penelitian

Kode	Uraian kebutuhan fungsional	Terkait masalah
KF-01	Tersedia cara yang terstruktur untuk mendefinisikan *user persona* eksplisit dan *user persona* implisit dalam bentuk skenario teks, sehingga variasi persona dapat dirancang secara konsisten dan digunakan kembali.	M-01
KF-02	Tersedia mekanisme untuk menjalankan pertanyaan yang sama pada beberapa *user persona* dan beberapa model bahasa, serta menyimpan keluaran model beserta informasi persona, model, dan jenis tugas yang digunakan.	M-02, M-04
KF-03	Tersedia format pencatatan hasil yang memungkinkan penilaian sederhana terhadap jawaban model, misalnya penandaan benar atau salah dan indikasi adanya *human bias*, sehingga hasil dapat dianalisis secara sistematis.	M-03, M-04

KF-01 berhubungan dengan kebutuhan untuk merepresentasikan persona secara eks-

plisit, sehingga skenario eksperimen dapat direplikasi. KF-02 menekankan pentingnya eksekusi skenario yang sama pada beberapa model dan persona dengan pencatatan hasil yang terstruktur. KF-03 memastikan bahwa keluaran model terdokumentasi dalam bentuk yang mendukung analisis kuantitatif maupun kualitatif tanpa menuntut skema penilaian yang terlalu kompleks.

III.2.3 Kebutuhan Nonfungsional

Kebutuhan nonfungsional berkaitan dengan kualitas pelaksanaan eksperimen, terutama dari sisi keterulangan, kesederhanaan implementasi, dan kemampuan pengembangan. Ringkasan kebutuhan nonfungsional ditunjukkan pada Tabel III.3.

Tabel III.3 Kebutuhan nonfungsional penelitian

Kode	Jenis kebutuhan	Uraian kebutuhan
KNF-01	Reproducibility	Proses eksperimen dapat diulang melalui skrip atau konfigurasi yang terdokumentasi, sehingga skenario persona, model, dan tugas dapat dijalankan kembali dengan pengaturan yang sama.
KNF-02	Simplicity	Implementasi eksperimen tetap sederhana dan dapat dijalankan dengan sumber daya komputasi yang wajar, misalnya melalui pemanggilan API tanpa memerlukan infrastruktur tambahan yang kompleks.
KNF-03	Extensibility	Rancangan eksperimen memungkinkan penambahan model atau *user persona* baru tanpa perubahan besar pada struktur keseluruhan, sehingga dapat menyesuaikan dengan ketersediaan model dan kebutuhan analisis lanjutan.

III.3 Analisis Pemilihan Solusi

Bagian ini membahas alternatif pendekatan yang dapat digunakan untuk melaksanakan eksperimen *multi model* dan *multi persona*, kemudian menjelaskan dasar pemilihan solusi yang digunakan dalam penelitian. Analisis dilakukan dengan mempertimbangkan kebutuhan struktur representasi *user persona*, konsistensi eksekusi lintas model dan lintas tugas, kemudahan pencatatan hasil untuk analisis, serta tingkat kerumitan implementasi.

III.3.1 Alternatif Solusi

Berdasarkan kebutuhan yang telah dirumuskan pada analisis kebutuhan, beberapa alternatif solusi yang dapat diidentifikasi adalah sebagai berikut.

1. Pendekatan evaluasi manual berbasis antarmuka percakapan. Pada alternatif ini, interaksi dengan *large language model* dilakukan langsung melalui antarmuka percakapan yang disediakan oleh penyedia layanan. *User persona* disisipkan ke dalam konteks, pertanyaan diajukan satu per satu, dan jawaban dicatat secara manual ke dalam dokumen atau lembar kerja. Setiap kombinasi model, persona, dan tugas dieksekusi secara terpisah. Pendekatan ini mudah dimulai karena tidak memerlukan pengembangan skrip, tetapi sangat bergantung pada prosedur manual dan kurang terstruktur ketika jumlah kombinasi skenario menjadi besar. Selain itu, reproduksi eksperimen menjadi bergantung pada kedisiplinan pencatatan dan rentan terhadap kesalahan manusia.
2. Skrip eksperimen semi terotomatisasi berbasis konfigurasi. Pada alternatif ini, definisi *user persona* eksplisit dan implisit, daftar model yang dievaluasi, serta kumpulan tugas dari *benchmark* seperti GSM8K dan MMLU-redux disimpan dalam berkas konfigurasi yang terstruktur. Skrip eksperimen membaca konfigurasi tersebut, membentuk *prompt* berdasarkan kombinasi model, persona, dan tugas, kemudian mengirim *prompt* ke model melalui antarmuka pemrograman aplikasi. Keluaran model, beserta metadata seperti nama model, jenis persona, jenis tugas, dan identitas soal, disimpan dalam berkas JSON pada direktori log. Tahap berikutnya, skrip analisis mengolah berkas JSON menjadi berkas CSV yang lebih ringkas untuk perhitungan metrik dan analisis lanjutan. Pendekatan ini menuntut pengembangan skrip, tetapi memberikan struktur yang jelas dan memudahkan pelaksanaan eksperimen berskala besar.
3. Kerangka evaluasi umum yang dapat digunakan kembali. Pada alternatif ini, dibangun sebuah kerangka evaluasi yang lebih umum, misalnya berupa pustaka atau layanan yang dirancang agar dapat digunakan kembali untuk berbagai studi terkait *user persona* pada *large language model*. Kerangka tersebut tidak hanya mencakup skrip eksekusi eksperimen berbasis konfigurasi, tetapi juga modul moduler untuk penjadwalan eksekusi, pengelolaan versi konfigurasi, penilaian otomatis, dan visualisasi hasil. Pendekatan ini berpotensi mendukung penggunaan jangka panjang dan kolaborasi yang lebih luas, namun memerlukan usaha perancangan dan implementasi yang lebih besar dibandingkan kebutuhan minimum untuk sebuah studi tugas akhir.

III.3.2 Analisis Penentuan Solusi

Penentuan solusi dilakukan dengan membandingkan ketiga alternatif berdasarkan beberapa kriteria utama, yaitu kemampuan merepresentasikan *user persona* dan skenario eksperimen secara terstruktur dan dapat digunakan kembali, konsistensi eksekusi lintas model dan lintas tugas, dukungan pencatatan hasil dan metadata untuk analisis kuantitatif dan kualitatif, keterulangan (*reproducibility*) proses eksperimen, serta tingkat kerumitan implementasi dan pemeliharaan. Ringkasan perbandingan alternatif ditunjukkan pada Tabel III.4, dengan skala kualitatif rendah, sedang, dan tinggi.

Tabel III.4 Perbandingan alternatif solusi

Kriteria	Evaluasi manual	Skrip semi terotomatisasi	Kerangka evaluasi umum
Representasi *user persona* dan skenario yang terstruktur	Rendah	Tinggi	Tinggi
Konsistensi eksekusi lintas model dan tugas	Rendah	Tinggi	Tinggi
Pencatatan hasil dan metadata untuk analisis	Rendah	Tinggi	Tinggi
Keterulangan (*reproducibility*) proses eksperimen	Rendah	Tinggi	Tinggi
Kerumitan implementasi dan pemeliharaan	Rendah	Sedang	Tinggi
Kemudahan penambahan model atau persona baru	Rendah	Tinggi	Tinggi

Pendekatan evaluasi manual relatif mudah digunakan pada tahap eksplorasi awal, tetapi tidak memadai untuk eksperimen *multi model* dan *multi persona* dengan jumlah kombinasi yang besar. Keterbatasan utama muncul pada konsistensi eksekusi, keterulangan eksperimen, serta pencatatan hasil yang sistematis.

Pendekatan kerangka evaluasi umum memberikan dukungan yang kuat terhadap struktur dan keterulangan, namun menuntut upaya perancangan arsitektur dan pengembangan perangkat lunak yang cukup besar. Beban tersebut berpotensi mengalihkan fokus dari tujuan utama penelitian, yaitu analisis empiris pengaruh *user persona* terhadap penalaran, kualitas jawaban, dan kecenderungan *human bias*.

Pendekatan skrip eksperimen semi terotomatisasi berbasis konfigurasi memberikan keseimbangan yang lebih sesuai. Representasi model, persona, dan tugas dapat diatur dalam direktori konfigurasi yang terpisah dari kode, sementara skrip eksekusi

dan analisis ditempatkan dalam direktori tersendiri. Keluaran eksperimen disimpan sebagai berkas JSON pada direktori log dan diolah lebih lanjut menjadi berkas CSV pada direktori hasil. Struktur ini mendukung konsistensi eksekusi, keterulangan eksperimen, dan analisis terukur tanpa memerlukan pembangunan kerangka evaluasi yang terlalu umum.

Berdasarkan pertimbangan tersebut, penelitian ini memilih pendekatan skrip eksperimen semi terotomatisasi berbasis konfigurasi sebagai solusi utama untuk melaksanakan eksperimen *multi model* dan *multi persona*.

BAB IV

DESAIN KONSEP SOLUSI

Bab ini memaparkan rancangan konsep solusi yang diusulkan untuk menjawab permasalahan yang telah dianalisis pada bab sebelumnya. Berdasarkan hasil analisis pemilihan solusi, pendekatan yang digunakan dalam penelitian ini adalah pengembangan sistem eksperimen terotomatisasi berbasis konfigurasi. Pembahasan dalam bab ini mencakup desain konseptual eksperimen, perancangan arsitektur perangkat lunak atau *evaluation pipeline*, serta spesifikasi implementasi data dan struktur berkas. Desain ini disusun untuk memenuhi kebutuhan fungsional terkait strukturisasi *user persona* dan konsistensi eksekusi lintas model.

IV.1 Desain Konseptual Eksperimen

Bagian ini menjelaskan dasar konseptual dari sistem eksperimen yang dikembangkan untuk mengkaji pengaruh *user persona* terhadap perilaku model bahasa. Desain konseptual ini berfungsi sebagai landasan arsitektur yang menghubungkan permasalahan metodologis pada Bab III dengan implementasi teknis yang dijelaskan pada Subbab IV.2 dan Subbab IV.3. Pendekatan yang digunakan menekankan pentingnya evaluasi yang terstruktur, terukur, dan bebas dari variasi input yang tidak relevan, sehingga setiap perubahan keluaran model dapat diinterpretasikan secara valid.

IV.1.1 Dekonstruksi Model Operasional Konvensional

Secara umum, praktik pengujian persona pada model bahasa masih banyak dilakukan menggunakan pendekatan manual atau semi-manual melalui *conversational interface*. Pada pendekatan ini, persona dituangkan langsung ke dalam *prompt* dan diberikan berulang-ulang setiap kali model diuji. Meskipun mudah diterapkan, pendekatan tersebut mengandung beberapa keterbatasan metodologis yang berdampak langsung pada validitas hasil eksperimen.

Keterbatasan pertama berkaitan dengan ketidakstabilan input atau *input framing variance*. Penelitian sebelumnya menunjukkan bahwa model bahasa sangat sensitif terhadap variasi kecil pada redaksi instruksi, termasuk perubahan tanda baca, panjang kalimat, atau urutan penyampaian (Turpin dkk. 2023; Zhou dkk. 2023). Dalam skenario pengujian manual, variasi ini sulit dikendalikan sepenuhnya sehingga dapat menimbulkan perbedaan respons yang bukan berasal dari persona, tetapi dari ketidakteraturan input.

Keterbatasan kedua adalah minimnya *observability*. Pengujian manual umumnya hanya menyimpan teks keluaran model, sedangkan informasi penting seperti *latency*, jumlah token, atau struktur penalaran tidak tercatat. Hal ini menyulitkan analisis mengenai bagaimana persona memengaruhi beban komputasi atau pola respons model, sebagaimana disorot oleh Naous et al. (Naous, Roziere, dkk. 2025).

Keterbatasan ketiga adalah aspek reproduktibilitas. Karena interaksi dilakukan melalui antarmuka percakapan, sulit untuk menjamin bahwa percobaan yang sama dapat dijalankan ulang dengan kondisi yang benar-benar identik. Hal ini bertentangan dengan prinsip penelitian empiris yang menuntut transparansi dan konsistensi prosedur.

IV.1.2 Konstruksi Model Sistem Terotomatisasi

Untuk mengatasi berbagai keterbatasan tersebut, penelitian ini mengusulkan model eksperimen terotomatisasi yang mengubah proses evaluasi dari pendekatan manual menjadi pendekatan berbasis data dan *automated orchestration*. Desain konseptual ini bertumpu pada tiga pilar utama.

Pilar pertama adalah *deterministic input configuration*. Setiap persona dan setiap pertanyaan benchmark diperlakukan sebagai objek data yang disimpan dalam format terstruktur. Pipeline menyusun instruksi secara programatik sehingga seluruh *byte-level input* konsisten untuk setiap iterasi. Dengan cara ini, variabel independen benar-benar terbatas hanya pada variasi persona.

Pilar kedua adalah peningkatan *data granularity*. Sistem merekam keluaran model beserta *telemetry* seperti *latency*, *token usage*, dan jejak penalaran apabila tersedia. Informasi ini memungkinkan analisis lebih komprehensif terhadap dampak persona, termasuk aspek efisiensi komputasi dan kecenderungan struktur respons.

Pilar ketiga adalah *scalable execution*. Mengingat jumlah kombinasi persona, model, dan butir soal yang besar, pipeline menerapkan eksekusi paralel dengan *asyn-*

chronous processing. Hal ini memungkinkan eksperimen diselesaikan dalam durasi yang lebih singkat tanpa mengorbankan konsistensi prosedural.

IV.1.3 Analisis Komparatif Metodologis

Transformasi dari pendekatan manual menuju pipeline terotomatisasi membawa implikasi metodologis yang signifikan. Tabel berikut merangkum perbedaan utama antara kedua pendekatan dari beberapa dimensi analisis penting.

Tabel IV.1 Perbandingan Validitas Metodologis antara Model Konvensional dan Model Terotomatisasi

Dimensi Analisis	**Pendekatan Konvensional**	**Pendekatan Terotomatisasi**
Input Control	Rentan terhadap variasi redaksi instruksi dan ketidakkonsistenan manual.	Instruksi dirakit secara programatik sehingga konsisten pada seluruh iterasi.
Data Granularity	Hanya menyimpan teks keluaran.	Merekam *latency*, token, dan metadata lainnya untuk analisis mendalam.
Storage Format	Tidak terstruktur dan sulit diproses ulang.	Menggunakan format JSON atau CSV yang siap dianalisis secara otomatis.
Reproducibility	Sulit menjamin kondisi eksperimen identik.	Seluruh parameter eksperimen terdokumentasi dan dapat diulang.
Scalability	Eksekusi linear dan memakan waktu.	Mendukung pemrosesan paralel berskala besar.

Melalui perancangan ini, sistem eksperimen yang dikembangkan mampu menghasilkan data yang lebih stabil, konsisten, dan transparan. Dengan demikian, setiap perbedaan performa model dapat ditelusuri kembali secara lebih jelas ke persona yang sedang diuji.

IV.2 Perancangan Arsitektur Perangkat Lunak (*Evaluation Pipeline*)

Subbab ini menjelaskan desain arsitektur perangkat lunak yang digunakan untuk merealisasikan *evaluation pipeline* sebagaimana dirumuskan pada Subbab IV.1. Arsitektur pipeline dirancang agar proses evaluasi dapat berjalan secara otomatis, konsisten, dan dapat direproduksi. Pendekatan ini memastikan bahwa setiap kombinasi

persona, model, dan *benchmark task* diuji dalam kondisi yang setara dan bebas dari variasi yang tidak diperlukan.

Pipeline yang dibangun bekerja sebagai rangkaian komponen yang saling berinteraksi: mulai dari pemuatan data, konstruksi instruksi, pengiriman permintaan ke model, hingga pencatatan *telemetry*. Seluruh proses tersebut bekerja dalam satu alur terintegrasi sehingga sistem mampu menangani jumlah evaluasi yang besar secara stabil.

IV.2.1 Arsitektur Alur Kerja Sistem

Secara garis besar, *evaluation pipeline* terbagi ke dalam empat komponen utama yang membentuk satu siklus pemrosesan yang berulang untuk setiap kombinasi persona dan butir soal. Keempat komponen tersebut adalah sebagai berikut.

1. *Configuration initialization and validation*.
   Tahap ini memuat seluruh konfigurasi sistem, definisi persona, dan *benchmark dataset* ke dalam memori. Validasi struktur data dilakukan untuk memastikan bahwa setiap persona memiliki *system instruction* yang lengkap dan setiap butir tugas memiliki pasangan pertanyaan dan jawaban acuan. Validasi awal ini penting untuk mencegah kesalahan format yang dapat menghentikan proses pada tahap berikutnya.
2. *Prompt construction engine*.
   Pada tahap ini, sistem membentuk dua jenis pesan: *system message* yang berisi identitas persona dan *user message* yang memuat pertanyaan dari benchmark. Penyusunan instruksi dilakukan menggunakan pola yang seragam untuk seluruh iterasi, sehingga setiap model menerima bentuk stimulus yang konsisten. Pendekatan ini menghilangkan variasi yang berasal dari perbedaan penulisan instruksi manual.
3. *Execution manager*.
   Komponen ini mengatur pengiriman permintaan ke model-model bahasa melalui *API interface*. Untuk mengatasi volume permintaan yang besar, *execution manager* menggunakan pendekatan eksekusi asinkron dengan *I/O concurrency*. Permintaan diatur dalam *task queue* dan dieksekusi dalam kelompok sesuai batas *rate limit*. Strategi ini mempercepat proses pengujian tanpa melampaui kapasitas layanan penyedia model.
4. *Telemetry logger*.
   Komponen terakhir bertanggung jawab menyimpan seluruh respons model dalam format terstruktur, termasuk *model output*, jumlah token, serta *latency*.

Data ini digunakan sebagai dasar analisis performa pada bab berikutnya. Dengan pembagian tersebut, pipeline dapat beroperasi secara modular namun tetap terpadu dalam satu alur pemrosesan.

IV.2.2 Algoritma Orkestrasi dan Konkurensi

Jumlah kombinasi persona, model, dan *benchmark tasks* menghasilkan volume permintaan yang sangat besar. Oleh karena itu, pipeline menerapkan mekanisme eksekusi asinkron untuk meningkatkan efisiensi pemrosesan.

Pipeline pertama-tama membangun sebuah *task queue* yang berisi seluruh pasangan persona–soal. Selanjutnya, *task queue* diproses dalam kelompok yang ukurannya ditentukan oleh kapasitas *concurrency*. Ketika satu kelompok tugas sedang diproses, sistem dapat menyiapkan kelompok berikutnya. Dengan demikian, waktu pemrosesan total dapat ditekan mendekati $O(N/C)$, di mana N adalah jumlah permintaan dan C adalah kapasitas konkurensi.

Apabila terjadi kegagalan seperti *timeout*, *connection reset*, atau batas *rate limit*, pipeline tidak langsung menghentikan seluruh proses. Sebaliknya, tugas tersebut dicatat dan dieksekusi ulang dengan *exponential backoff*. Pendekatan ini membuat pipeline tetap stabil meskipun dijalankan dalam waktu yang panjang.

IV.2.3 Mekanisme Injeksi Konteks Persona

Mekanisme injeksi persona merupakan elemen penting untuk memastikan bahwa pengaruh persona dapat diukur dengan jelas. Pipeline menerapkan dua tahap injeksi konteks yang bersifat tetap dan hanya dilakukan satu kali untuk setiap persona sebelum evaluasi dimulai.

Tahap pertama adalah *persona context initialization*. Pada tahap ini, sistem menyusun pesan awal yang merangkum identitas dan karakter persona. Pesan ini berfungsi membangun *cognitive framing* awal pada model, baik untuk persona eksplisit maupun implisit. Tahap ini memastikan bahwa model berada dalam kondisi persona yang konsisten sebelum diberikan tugas.

Tahap kedua adalah *persona warm-up message*. Pesan ini digunakan untuk memastikan bahwa model memberikan respons yang sesuai dengan identitas persona. Respons dari tahap ini tidak digunakan dalam evaluasi, tetapi berfungsi sebagai verifikasi bahwa proses injeksi berhasil.

Setelah kedua tahap ini selesai, pipeline tidak lagi mengulangi injeksi persona untuk setiap pertanyaan. Identitas yang telah ditanamkan pada awal percakapan tetap digunakan selama seluruh rangkaian pengujian. Model kemudian langsung memproses seluruh soal pada GSM8K dan MMLU-Redux dalam kondisi persona yang sama. Pendekatan ini memastikan bahwa variasi keluaran model berasal dari perbedaan persona, bukan dari perbedaan struktur instruksi.

IV.2.4 Mekanisme Toleransi Kesalahan dan Persistensi Status

Pipeline dirancang agar tetap stabil meskipun menghadapi gangguan selama proses pengujian. Dua mekanisme utama digunakan untuk menjamin integritas data dan keberlanjutan proses.

Pertama, sistem menerapkan *state persistence*. Setelah setiap tugas berhasil diproses, status kemajuan dicatat sehingga apabila terjadi interupsi, pipeline dapat dilanjutkan kembali tanpa mengulangi tugas yang sudah selesai.

Kedua, gangguan sementara ditangani dengan *error handling* berbasis penjadwalan ulang adaptif. Tugas yang gagal tidak langsung dihentikan, tetapi dijalankan kembali setelah jeda waktu tertentu. Dengan kombinasi kedua strategi ini, pipeline dapat menyelesaikan seluruh rangkaian evaluasi meskipun terjadi kendala jaringan atau batasan layanan eksternal.

IV.3 Implementasi Data, Struktur Berkas, dan Keluaran Pipeline

Subbab ini menjelaskan bagaimana pipeline yang telah dirancang pada bagian sebelumnya terealisasi dalam bentuk struktur data, organisasi direktori, serta keluaran eksperimen yang dihasilkan. Implementasi ini berperan penting dalam memastikan bahwa seluruh proses pemuatan data, injeksi konteks, inferensi, pencatatan log, dan agregasi hasil berlangsung secara terstruktur, dapat ditelusuri, serta mendukung keterulangan eksperimen.

IV.3.1 Organisasi Direktori dan Artefak Data

Struktur direktori disusun secara hierarkis untuk memisahkan fungsi-fungsi inti dalam pipeline. Empat komponen utama yang digunakan adalah sebagai berikut.

1. *Root directory*. Berisi skrip pemanggil pipeline serta utilitas eksekusi.
2. *Configuration directory*. Menyimpan konfigurasi model, kredensial layanan, dan parameter eksekusi.

3. *Input assets directory*. Berisi definisi persona dan *benchmark datasets* yang telah dinormalisasi.
4. *Results directory*. Memuat log inferensi per butir soal dan tabel agregasi hasil dalam format terstruktur.

Struktur ini memudahkan proses audit dan memastikan seluruh artefak eksperimen terdokumentasi dengan baik.

IV.3.2 Subsistem Perangkat Lunak dan Alur Transformasi Data

Pipeline diimplementasikan melalui empat subsistem perangkat lunak yang berinteraksi secara berurutan:

1. *Execution orchestration subsystem*. Bertugas membentuk *task queue* yang memuat seluruh kombinasi model, persona, dan pertanyaan.
2. *Model communication subsystem*. Mengelola pembentukan pesan, pengiriman *prompt*, penanganan respons, serta batas layanan.
3. *Monitoring subsystem*. Menyediakan mekanisme *checkpointing* sehingga eksekusi dapat dilanjutkan setelah gangguan.
4. *Analysis subsystem*. Mengolah log granular menjadi tabel hasil, termasuk perhitungan akurasi, penggunaan token, *latency*, dan agregasi antar-persona.

Transformasi data berlangsung dari log mentah → granular CSV → agregasi CSV → tabel ringkasan.

IV.3.3 Representasi Persona dan Injeksi Konteks

Persona direpresentasikan dalam format terstruktur yang berisi identitas dan narasi. Representasi tersebut dikonversi menjadi *system instruction* yang digunakan pada tahap injeksi konteks.

Injeksi konteks terdiri atas dua tahap tetap yang dilakukan satu kali untuk setiap persona:

1. *Persona grounding*, yaitu pengenalan identitas dan gaya naratif ke dalam konteks model.
2. *Warm-up interaction*, yaitu satu interaksi pemanasan untuk memastikan model merespons sesuai karakter persona.

Setelah kedua tahap ini, pipeline mengirim seluruh pertanyaan GSM8K dan MMLU-Redux tanpa mengulang injeksi persona. Pendekatan ini memastikan bahwa pengaruh persona tetap konsisten sepanjang keseluruhan sesi.

IV.3.4 Contoh Log Inferensi

Pipeline mencatat hasil inferensi dalam format terstruktur yang berisi metadata eksekusi, jawaban akhir, serta *telemetry* penggunaan token dan latensi. Dua contoh berikut menunjukkan keluaran model tanpa *reasoning trace* dan dengan *reasoning trace*.

Log tanpa *reasoning trace*


```
{
  "run": {"model_id": "example-model", "question_id": "gsm8k_00001"},
  "response": {
    "choices": [{"message": {"content": "Let's break down the problem..."}}],
    "usage": {"prompt_tokens": 211, "completion_tokens": 197}
  }
}
```


Gambar IV.1 Contoh log inferensi tanpa *reasoning trace*

Log dengan *reasoning trace*


```
{
  "run": {"model_id": "example-model-reasoning", "question_id": "gsm8k_00003"}
  "response": {
    "choices": [{
       "message": {
         "content": "Final answer: 70000",
         "reasoning": "First compute the purchase cost..."
       }}],
    "usage": {"completion_tokens": 867, "reasoning_tokens": 485}
  }
}
```


Gambar IV.2 Contoh log inferensi dengan *reasoning trace*

IV.3.5 Ringkasan Hasil Eksperimen (Full Table)

Ringkasan hasil eksperimen GSM8K untuk keseluruhan model dan persona ditampilkan pada Tabel IV.2. Tabel ini menunjukkan perbandingan performa lintas model pada metrik akurasi, penggunaan token, dan *latency*.

Tabel IV.2 Ringkasan Hasil Eksperimen GSM8K untuk Seluruh Model dan Persona

Model	Persona	Total Q	Correct	Accuracy (%)	Total Prompt T.	Total Compl. T.	Total Tokens	Total Reas. T.	A
Bert Nebulon Alpha	man_implicits	610	593	97.21	125551	159699	285250	0	
Bert Nebulon Alpha	woman_implicits	641	627	97.26	17138	163670	335208	0	
Grok 4.1 Fast	man_implicits	1315	1242	94.45	461404	863825	1325229	634958	
Grok 4.1 Fast	woman_implicits	1316	1254	95.36	538871	844065	1422736	671910	
Nvidia Nemotron-nano-12B-v2-VL	man_implicits	1305	1224	93.79	268984	887065	1156049	0	
Nvidia Nemotron-nano-12B-v2-VL	woman_implicits	1315	1248	94.98	356096	1634188	1986284	0	

IV.3.6 Contoh Ringkasan Satu Model

Untuk mempermudah interpretasi, Tabel berikut menampilkan contoh ringkasan singkat untuk satu model, yaitu Grok 4.1 Fast.

Tabel IV.3 Contoh Ringkasan Hasil untuk Model Grok 4.1 Fast

Persona	Total Q	Correct	Accuracy (%)	Avg Tokens	Avg Latency (s)
man_implicits	1315	1242	94.45	100755/1315	9.591
woman_implicits	1316	1254	95.36	108132/1316	9.487

Kedua tabel ini menjadi dasar untuk analisis performa dan karakteristik respons model yang dibahas lebih lanjut pada Bab V.

BAB V

RENCANA SELANJUTNYA

Jelaskan secara detail langkah-langkah rencana selanjutnya, hal-hal yang diperlukan atau akan disiapkan, dan risiko dan mitigasinya, yang meliputi:

1. Rencana implementasi, termasuk alat dan bahan yang diperlukan, lingkungan, konfigurasi, biaya, dan sebagainya.
2. Desain pengujian dan evaluasi, misalnya metode verifikasi dan validasi.
3. Analisis risiko dan mitigasi, misalnya tindakan selanjutnya jika ada yang tidak berjalan sesuai rencana.

DAFTAR PUSTAKA

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, dkk. 2021. "On the Opportunities and Risks of Foundation Models". *arXiv preprint arXiv:2108.07258,* https://arxiv.org/abs/2108.07258.

Cobbe, Karl, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, dkk. 2021. "Training Verifiers to Solve Math Word Problems". *arXiv preprint arXiv:2110.14168,* https://arxiv.org/abs/2110.14168.

Edinburgh Dataset Analytics Working Group. 2024. *MMLU-Redux 2.0 Dataset.* https://huggingface.co/datasets/edinburgh-dawg/mmlu-redux-2.0. Versi kurasi ulang MMLU dengan 57 subjek dan 100 butir soal per subjek.

Gema, Aryo Pradipta, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, dkk. 2024. "Are We Done with MMLU?" *arXiv preprint arXiv:2406.04127,* https://arxiv.org/abs/2406.04127.

Gupta, Shashank, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, dan Tushar Khot. 2024. "Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned Language Models". Dalam *Proceedings of the Twelfth International Conference on Learning Representations.* https://openreview.net/forum?id=kGteeZ18Ir.

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, dan Jacob Steinhardt. 2021. "Measuring Massive Multitask Language Understanding". *International Conference on Learning Representations,* https://arxiv.org/abs/2009.03300.

Naous, Tarek, Baptiste Roziere, dkk. 2025. "Training and Evaluating User Language Models". *arXiv preprint arXiv:2510.06552,* https://arxiv.org/abs/2510.06552.

Sap, Maarten, Hannah Rashkin, Derek Chen, Ronan Le Bras, dan Yejin Choi. 2019. "SocialIQA: Commonsense Reasoning about Social Interactions". Dalam *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing.* Hong Kong, China. https://arxiv.org/abs/1904.09728.

Tseng, Yu-Min, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, dan Yun-Nung Chen. 2024. "Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization". Dalam *Findings of the Association for Computational Linguistics: EMNLP 2024,* 16612–16631. https://aclanthology.org/2024.findings-emnlp.969.

Turpin, Miles, dkk. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Reasoning". *arXiv preprint arXiv:2305.04388,* https://arxiv.org/abs/2305.04388.

Weidinger, Laura, John Mellor, Maribeth Rauh, Christopher Griffin, Iason Gabriel, Jonathan Uesato, Po-Sen Huang, Zachary Kenton, Tom B. Brown, dkk. 2021. "Ethical and Social Risks of Harm from Language Models". *arXiv preprint arXiv:2112.04359,* https://arxiv.org/abs/2112.04359.

Zhao, Yanhao, Eric Wallace, Shi Feng, Mohit Singh, dan Matt Gardner. 2021. "Calibrate Before Use: Improving Few-Shot Performance of Language Models". Dalam *Proceedings of the International Conference on Machine Learning,* 12697–12706.

Zhou, Luozhi, dkk. 2023. "Large Language Models Are Sensitive to Prompt Framing". *arXiv preprint arXiv:2310.05400,* https://arxiv.org/abs/2310.05400.