# 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** . . . . . **vi**

**DAFTAR TABEL** . . . . . **vii**

**DAFTAR KODE** . . . . . **viii**

**I PENDAHULUAN** . . . . . **1**
I.1 Latar Belakang . . . . . 1
I.2 Rumusan Masalah . . . . . 3
I.3 Tujuan Penelitian . . . . . 3
I.4 Batasan Masalah . . . . . 4
I.5 Metodologi . . . . . 4
I.5.1 Tahap 1: Investigasi Awal dan Pengumpulan Fakta . . . . . 5
I.5.2 Tahap 2: Pencarian, Pengelompokan, dan Penapisan Literatur . . . . 5

**II STUDI LITERATUR** . . . . . **7**
II.1 Large Language Model . . . . . 7
II.1.1 Konsep dan Karakteristik Dasar . . . . . 7
II.1.2 Representasi Bahasa dan Pemahaman Instruksi . . . . . 8
II.1.3 Penalaran dan Dinamika Perilaku Model . . . . . 8
II.1.4 Dimensi Sosial dalam Pemrosesan Bahasa . . . . . 9
II.2 Persona dalam Interaksi Model Bahasa . . . . . 10
II.2.1 Definisi dan Ruang Lingkup Persona . . . . . 10
II.2.2 Persona Eksplisit dan Persona Implisit . . . . . 10
II.2.3 Peran Persona dalam Interaksi dengan LLM . . . . . 11
II.3 Pengaruh Persona terhadap Perilaku LLM . . . . . 12
II.3.1 Pengaruh Persona terhadap Penalaran Model . . . . . 12
II.3.2 Pengaruh Persona terhadap Gaya dan Struktur Respons . . . . . 13
II.3.3 Faktor yang Memperkuat Efek Persona . . . . . 13
II.4 Bias dalam Respons LLM . . . . . 14
II.4.1 Bentuk-bentuk Bias pada Model Bahasa . . . . . 14
II.4.2 Konsekuensi Bias terhadap Keluaran Model . . . . . 15
II.4.3 Kaitannya dengan Variasi Persona . . . . . 16
II.5 Evaluasi Penalaran dan Benchmark . . . . . 16
II.5.1 Benchmark Penalaran dan Pengetahuan . . . . . 17
II.5.2 Benchmark Sosial dan Moral . . . . . 17

II.5.3 Tantangan Evaluasi Berbasis Persona . . . . . 18
II.6 Penelitian Terdahulu dan Kesenjangan Penelitian . . . . . 18
II.6.1 Ringkasan Literatur Terkait . . . . . 19
II.6.2 Keterbatasan Penelitian Sebelumnya . . . . . 20
II.6.3 Posisi dan Kontribusi Penelitian Ini . . . . . 20

**III ANALISIS MASALAH . . . . . 22**
III.1 Analisis Kondisi Saat Ini . . . . . 22
III.2 Analisis Kebutuhan . . . . . 25
III.2.1 Identifikasi Masalah Pengguna . . . . . 25
III.2.2 Kebutuhan Fungsional . . . . . 26
III.2.3 Kebutuhan Nonfungsional . . . . . 27
III.3 Analisis Pemilihan Solusi . . . . . 27
III.3.1 Alternatif Solusi . . . . . 28
III.3.2 Analisis Penentuan Solusi . . . . . 29

**IV DESAIN KONSEP SOLUSI . . . . . 31**
IV.1 Desain Konseptual Eksperimen . . . . . 31
IV.1.1 Dekonstruksi Model Operasional Konvensional (*Existing Model*) . . 31
IV.1.2 Konstruksi Model Sistem Terotomatisasi (*Proposed Model*) . . . . . 32
IV.1.3 Analisis Komparatif Metodologis . . . . . 33
IV.2 Perancangan Arsitektur Perangkat Lunak (*Evaluation Pipeline*) . . . . . 35
IV.2.1 Arsitektur Alur Kerja Sistem . . . . . 35
IV.2.2 Algoritma Orkestrasi dan Konkurensi . . . . . 36
IV.2.3 Spesifikasi Mekanisme Injeksi Konteks . . . . . 37
IV.2.4 Mekanisme Toleransi Kesalahan dan Persistensi Status . . . . . 38
IV.3 Implementasi Data dan Struktur Berkas . . . . . 38
IV.3.1 Organisasi Direktori Proyek . . . . . 39
IV.3.2 Implementasi Modul Perangkat Lunak . . . . . 39
IV.3.3 Spesifikasi Artefak Data . . . . . 40
IV.3.4 Ilustrasi Berkas Data Eksperimen . . . . . 41
IV.3.4.1 Contoh Konfigurasi Persona . . . . . 41
IV.3.4.2 Contoh Log Keluaran GSM8K . . . . . 41
IV.4 Rancangan Evaluasi dan Metrik . . . . . 44
IV.4.1 Metrik Performansi Penalaran . . . . . 44
IV.4.2 Metrik Efisiensi Komputasi . . . . . 44
IV.4.3 Format Data Analisis . . . . . 45
IV.4.4 Ilustrasi Data Hasil Eksperimen . . . . . 45

**V RENCANA SELANJUTNYA . . . . . 47**

# DAFTAR GAMBAR

IV.3 Definisi Persona Implisit pada `persona_echo.json` . . . . . 41
IV.4 Struktur Injeksi Konteks pada `persona_warmup.json` . . . . . 42
IV.5 Contoh Log Eksekusi `gsm8k_00001.json` . . . . . 42
IV.6 Contoh Log Eksekusi `gsm8k_00003.json` dengan *Reasoning Trace* 43

# DAFTAR TABEL

III.1 Daftar masalah penelitian terkait *user persona* pada LLM . . . . . 24
III.2 Kebutuhan fungsional penelitian . . . . . 26
III.3 Kebutuhan nonfungsional penelitian . . . . . 27
III.4 Perbandingan alternatif solusi . . . . . 29

IV.1 Analisis Komparatif Validitas Metodologis . . . . . 34
IV.2 Sampel Data Hasil Granular (Grok 4.1) . . . . . 46
IV.3 Sampel Data Hasil Teragregasi (Ringkasan) . . . . . 46

# 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

Perancangan arsitektur eksperimen pada penelitian ini didasarkan pada urgensi untuk mentransformasi mekanisme evaluasi *Large Language Model* (LLM) dari pendekatan eksploratif yang bersifat *ad-hoc* menjadi suatu sistem orkestrasi yang deterministik dan terukur. Subbab ini menguraikan model konseptual dari sistem yang dirancang serta menyajikan analisis komparatif kritis terhadap validitas metodologis antara pendekatan konvensional dan pendekatan terotomatisasi yang diusulkan. Fokus analisis diletakkan pada aspek integritas pengendalian variabel input serta granularitas akuisisi data output.

### IV.1.1 Dekonstruksi Model Operasional Konvensional (*Existing Model*)

Berdasarkan analisis kondisi saat ini, lanskap evaluasi pengaruh atribut pengguna terhadap model bahasa masih didominasi oleh pendekatan operasional yang mengandalkan eksekusi manual atau semi-otomatis. Dalam model ini, interaksi dengan model dilakukan melalui antarmuka percakapan atau *conversational interface* di mana *persona* disisipkan sebagai bagian dari teks instruksi atau *prompt* secara manual untuk setiap sesi pengujian.

Dari perspektif riset yang ketat, model operasional ini mengandung dua cacat metodologis fundamental yang mengancam validitas internal eksperimen. Kelemahan pertama berkaitan dengan instabilitas variabel kontrol atau *input noise*. Validitas eksperimen sangat bergantung pada kemampuan peneliti mengisolasi variabel independen, yaitu variasi *user persona*. Namun, literatur menunjukkan bahwa LLM memiliki sensitivitas ekstrem terhadap variasi sintaksis. Turpin et al. (Turpin dkk. 2023) membuktikan bahwa perubahan minor pada *prompt framing* dapat memicu deviasi signifikan pada rantai penalaran model atau *Chain-of-Thought*. Dalam metode manual, inkonsistensi pengetikan instruksi seperti perbedaan spasi, tanda baca, atau pemilihan diksi merupakan variabel pengganggu atau *confounding variable* yang sulit dihindari. Kondisi ini berisiko menimbulkan bias eksperimental karena perubahan respons model tidak sepenuhnya disebabkan oleh karakteristik *persona*, melainkan oleh artefak input yang tidak disengaja.

Kelemahan kedua terletak pada defisit observabilitas atau *data granularity*. Metode konvensional memiliki keterbatasan dalam merekam telemetri komputasi internal model karena umumnya hanya menangkap teks jawaban akhir atau *completion text* yang disalin ke lembar kerja manual. Metadata krusial seperti latensi inferensi dan jumlah token sering kali luput dari pencatatan. Akibatnya, peneliti kehilangan visibilitas terhadap beban kognitif model, seperti apakah *persona* tertentu menyebabkan model membutuhkan waktu komputasi lebih lama atau menghasilkan uraian yang lebih panjang. Padahal, metrik tersebut merupakan indikator penting dalam analisis bias dan evaluasi model pengguna sebagaimana disarankan oleh Naous et al. (Naous, Roziere, dkk. 2025).

Gambar IV.1 Model Konseptual Sistem Eksperimen Konvensional

### IV.1.2 Konstruksi Model Sistem Terotomatisasi (*Proposed Model*)

Guna memitigasi defisiensi metodologis tersebut, penelitian ini merekayasa model sistem eksperimen yang sepenuhnya terotomatisasi atau *fully automated pipeline* dengan arsitektur berbasis data. Paradigma ini menggeser fokus penelitian dari sekadar interaksi percakapan menjadi orkestrasi data yang presisi. Desain konseptual sistem ini dibangun di atas tiga pilar utama yang menjamin rigoritas ilmiah.

Pilar pertama adalah penerapan determinisme input berbasis konfigurasi. Sistem ini meniadakan variabilitas input manusia dengan memperlakukan *persona* sebagai objek data statis yang didefinisikan dalam berkas konfigurasi berekstensi JSON. Sistem melakukan injeksi konteks secara terprogram atau *programmatic injection*

untuk memastikan bahwa setiap model menerima stimulus yang identik secara *byte-level*. Pendekatan ini secara efektif mengeliminasi *framing bias* yang tidak diinginkan, sehingga setiap variasi pada keluaran dapat diasosiasikan secara kausal dengan variabel *persona* yang sedang diuji.

Pilar kedua adalah akuisisi telemetri multi-dimensi. Berbeda dengan pencatatan manual yang hanya bersifat tekstual, sistem usulan dirancang untuk menangkap spektrum data yang lengkap. Respons model disimpan sebagai objek data terstruktur yang memuat dimensi linguistik berupa teks jawaban dan jejak penalaran atau *reasoning trace* untuk keperluan analisis kualitatif. Selain itu, sistem menangkap dimensi komputasional berupa metadata penggunaan token dan durasi eksekusi. Data ini memungkinkan analisis korelasi antara *persona* dengan efisiensi penalaran model. Sistem juga merekam dimensi konfigurasi berupa parameter model yang digunakan saat eksekusi guna menjamin aspek *reproducibility*.

Pilar ketiga adalah skalabilitas eksekusi paralel. Mengingat kebutuhan untuk mengevaluasi matriks kombinasi yang kompleks antara berbagai model, *persona* baik eksplisit maupun implisit, dan tugas penalaran, sistem dirancang menggunakan arsitektur pemrosesan asinkron. Hal ini memungkinkan sistem mengirimkan permintaan ke berbagai API model secara serentak untuk mengatasi kendala waktu yang melekat pada metode serial konvensional.

Gambar IV.2 Model Konseptual Sistem Eksperimen Terotomatisasi

### IV.1.3 Analisis Komparatif Metodologis

Transformasi dari model konseptual *Existing* ke *Proposed* bukan sekadar peningkatan efisiensi operasional, melainkan peningkatan validitas penelitian. Tabel IV.1 menyajikan analisis komparatif mendalam mengenai implikasi ilmiah dari kedua pendekatan tersebut yang ditinjau dari dimensi pengendalian variabel, integritas data, dan reproduktibilitas.

Melalui desain ini, penelitian memastikan bahwa setiap kesimpulan mengenai bias atau perubahan penalaran yang ditarik nantinya didasarkan pada data yang berintegritas tinggi, lengkap, dan diperoleh melalui prosedur yang dapat dipertanggungjawabkan secara ilmiah.

Tabel IV.1 Analisis Komparatif Validitas Metodologis

| **Dimensi Analisis** | **Sistem Konvensional (*Existing*)** | **Sistem Usulan (*Proposed*)** |
|---|---|---|
| *Pengendalian Variabel* | *Stokastik.* Rentan terhadap gangguan input manual yang dapat mendistorsi kausalitas efek *persona* akibat sensitivitas *framing*. | *Deterministik.* Konfigurasi statis dan injeksi otomatis menjamin isolasi variabel independen yang presisi dan konsisten. |
| *Granularitas Data* | *Dangkal (Tekstual).* Hanya menangkap hasil akhir sehingga kehilangan nuansa proses internal model seperti latensi dan efisiensi token. | *Dalam (Telemetri).* Menangkap metadata performa yang mengindikasikan beban kognitif dari adopsi *persona* secara granular. |
| *Format Penyimpanan* | *Tidak Terstruktur.* Teks mentah dalam lembar kerja yang sulit diproses ulang secara komputasi dan rentan kesalahan salin. | *Terstruktur (JSON/CSV).* Data siap olah yang mendukung analisis statistik otomatis dan deteksi pola bias yang sistematis. |
| *Reproduktibilitas* | *Rendah.* Parameter eksperimen sering kali tidak terdokumentasi dengan baik sehingga menyulitkan verifikasi pihak ketiga. | *Tinggi.* Seluruh kondisi eksperimen terenkapsulasi dalam kode sumber dan berkas konfigurasi yang dapat diaudit dan dijalankan ulang. |
| *Cakupan Eksperimen* | *Terbatas.* Kendala waktu eksekusi linear membatasi jumlah kombinasi model dan *persona* yang dapat diuji secara layak. | *Masif.* Mendukung evaluasi skala besar untuk memetakan perilaku model pada spektrum *persona* yang luas. |

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

Realisasi dari desain konseptual sistem eksperimen diwujudkan melalui pengembangan *Evaluation Pipeline*, sebuah kerangka kerja perangkat lunak yang berfungsi sebagai orkestrator eksekusi pengujian. Subbab ini menjabarkan spesifikasi teknis dari alur kerja sistem, algoritma orkestrasi, mekanisme manipulasi data instruksi, serta strategi manajemen ketahanan sistem. Perancangan ini difokuskan untuk memenuhi kebutuhan fungsional terkait otomatisasi pengujian serta kebutuhan non-fungsional terkait efisiensi waktu, skalabilitas, dan integritas pencatatan data.

### IV.2.1 Arsitektur Alur Kerja Sistem

Secara fungsional, arsitektur perangkat lunak dirancang menggunakan pendekatan modular yang memisahkan logika pemrosesan data atau *data processing* dari logika komunikasi jaringan atau *network communication*. Alur kerja sistem dibagi menjadi empat komponen sekuensial yang saling berinteraksi untuk membentuk satu siklus eksperimen yang utuh.

1. *Inisialisasi dan validasi konfigurasi.*
   Komponen ini bertindak sebagai gerbang awal yang bertanggung jawab memuat seluruh aset data statis ke dalam memori. Sistem membaca berkas definisi *persona* dan *dataset* tugas penalaran, kemudian melakukan validasi skema data secara ketat. Proses validasi tersebut memastikan bahwa setiap objek *persona* memiliki atribut instruksi yang tidak kosong dan setiap butir soal memiliki struktur pertanyaan serta kunci jawaban yang valid. Mekanisme deteksi dini ini diterapkan guna mencegah kegagalan proses di tengah eksekusi yang berpotensi membuang sumber daya komputasi.
2. *Mesin konstruksi instruksi atau prompt engine.*
   Unit pemrosesan ini berfungsi mentransformasi data mentah menjadi objek pesan yang siap dikirimkan ke model. Penggabungan string dilakukan antara atribut *persona* dan atribut pertanyaan berdasarkan templat pesan standar. Pada tahap ini, parameter operasional yang bersifat statis, seperti batas maksimum token keluaran, turut disematkan untuk menjamin bahwa kondisi eksperimen tetap terkendali dan konsisten di seluruh iterasi pengujian.
3. *Pengelola eksekusi atau execution dispatcher.*
   Tanggung jawab utama dari subsistem ini adalah mengelola komunikasi dengan antarmuka pemrograman aplikasi (API) dari berbagai penyedia model bahasa. Mengingat volume permintaan yang masif, manajemen antrean tugas diterapkan untuk mengatur distribusi muatan pesan ke jaringan secara efisien

tanpa membebani *bandwidth*.

4. *Pencatat telemetri atau telemetry logger.*
   Berbeda dengan metode pencatatan konvensional yang hanya menyimpan teks luaran, komponen pelaporan ini dirancang untuk menangkap aliran data respons secara utuh. Metadata teknis, meliputi durasi latensi eksekusi dan statistik penggunaan token, diekstrak dan disimpan ke dalam sistem berkas lokal secara *real-time*. Pendekatan penulisan langsung ini diadopsi untuk memitigasi risiko kehilangan data apabila terjadi terminasi program secara tidak terduga.

### IV.2.2 Algoritma Orkestrasi dan Konkurensi

Tantangan skalabilitas dalam eksperimen ini diatasi melalui implementasi algoritma eksekusi asinkron atau *asynchronous execution*. Dalam paradigma pemrograman sekuensial tradisional, sistem harus menunggu satu permintaan selesai diproses sebelum mengirimkan permintaan berikutnya, yang mengakibatkan akumulasi waktu tunggu atau *latency* yang besar. Sebagai solusi, pendekatan konkurensi I/O atau *I/O Concurrency* diterapkan untuk mengoptimalkan penggunaan waktu.

Untuk memperjelas logika orkestrasi tersebut, prosedur eksekusi eksperimen didefinisikan secara formal melalui Algoritma 4.1 berikut. Algoritma ini menjabarkan bagaimana penanganan konkurensi dilakukan untuk memproses himpunan *persona* ($P$) dan himpunan tugas ($T$) secara efisien dengan batasan *rate limit* ($C$).

```
Algoritma 4.1: Prosedur Eksekusi Eksperimen Paralel

Input : Himpunan Persona P, Himpunan Tugas T, Batas Konkurensi C
Output: Himpunan Log L

Function RunExperiment(P, T):
  1. Inisialisasi Antrean Tugas Q <- Kosong
  2. Untuk setiap p dalam P lakukan:
       Untuk setiap t dalam T lakukan:
         Prompt <- ConstructPrompt(p.instruction, t.question)
         Enqueue(Q, Prompt)

  3. Inisialisasi Semaphore S dengan kapasitas C
  4. While Q tidak kosong lakukan secara Asinkron:
       Batch <- DequeueBatch(Q, C)
```

```
    Untuk setiap item i dalam Batch lakukan secara Paralel:
      Acquire(S)
      Coba:
        Respons <- AsyncCallAPI(i.prompt, i.config)
        Metadata <- ExtractTelemetry(Respons)
        SaveLog(Respons, Metadata)
        Tambahkan ke L
      Tangkap Galat:
        LogGalat(i)
        RetryWithBackoff(i)
      Akhirnya:
        Release(S)
5. Return L
```

Algoritma di atas menunjukkan bahwa eksekusi sekuensial dengan kompleksitas waktu $O(N)$ telah digantikan dengan pemanfaatan *semaphore* untuk mengelola konkurensi, sehingga kompleksitas waktu eksekusi dapat ditekan mendekati $O(N/C)$ di mana $C$ adalah kapasitas *throughput* API.

### IV.2.3 Spesifikasi Mekanisme Injeksi Konteks

Integritas validitas internal eksperimen sangat bergantung pada kemampuan sistem dalam mengisolasi variabel independen, yaitu *persona*, dari variabel tugas. Untuk mencapai isolasi ini, diterapkan spesifikasi injeksi konteks berbasis peran atau *role-based injection* yang memanfaatkan struktur protokol pesan pada *Large Language Model* modern.

Definisi *persona* dipetakan ke dalam segmen *System Message*. Segmen ini berfungsi sebagai instruksi tingkat tinggi yang mendefinisikan identitas, nada bicara, dan batasan perilaku model. Dengan menempatkan *persona* pada posisi ini, kondisi kognitif model secara efektif dikunci sebelum memproses informasi lainnya. Isi dari segmen tersebut bersifat statis untuk satu varian *persona* tertentu, menjamin bahwa *framing* identitas tidak berubah sepanjang sesi eksperimen.

Sebaliknya, materi uji dari *benchmark* seperti GSM8K atau MMLU ditempatkan pada segmen *User Message*. Segmen ini diperlakukan sebagai stimulus eksternal yang harus direspons oleh model sesuai dengan identitas yang telah ditanamkan sebelumnya. Pemisahan semantik antara *System* dan *User* ini mencegah terjadinya kebocoran konteks atau *context leakage* di mana instruksi tugas bercampur aduk

dengan instruksi identitas, sehingga memungkinkan penarikan kesimpulan kausal yang lebih kuat mengenai pengaruh *persona* terhadap performa penalaran.

### IV.2.4 Mekanisme Toleransi Kesalahan dan Persistensi Status

Mengingat durasi eksperimen yang panjang dan ketergantungan pada layanan jaringan eksternal, penerapan arsitektur tahan kegagalan atau *fault-tolerant architecture* menjadi elemen krusial untuk menjamin keberhasilan pengumpulan data. Implementasi mekanisme ini didasarkan pada dua pilar utama, yaitu persistensi status atau *state persistence* dan pemulihan otomatis atau *automated recovery*.

1. *Mekanisme checkpointing.*
   Sebuah subsistem pemantau atau *checkpoint monitor* diimplementasikan untuk menyimpan status eksekusi ke dalam penyimpanan lokal secara periodik. Setiap kali sebuah tugas berhasil diselesaikan dan log-nya tersimpan, indeks penanda atau *cursor* pada berkas pelacakan diperbarui. Hal ini menjamin sifat *idempotency* pada sistem, di mana jika proses terhenti akibat kegagalan daya atau gangguan jaringan fatal, eksekusi ulang sistem tidak akan menduplikasi permintaan yang sudah berhasil, melainkan secara cerdas melanjutkan proses atau *resume* dari indeks tugas terakhir yang belum selesai.
2. *Strategi penanganan galat transien.*
   Untuk menangani kegagalan jaringan yang bersifat sementara atau *transient errors*, seperti *timeout* atau kode status HTTP 429 yang menandakan *Too Many Requests*, strategi *Exponential Backoff* diterapkan. Ketika galat terdeteksi, proses tidak langsung dihentikan, melainkan dilakukan penundaan eksekusi dengan durasi yang meningkat secara eksponensial ($t = base \times 2^n$) sebelum mencoba mengirimkan ulang permintaan. Mekanisme ini mencegah pembebanan berlebih pada server API sekaligus meningkatkan probabilitas keberhasilan permintaan pada percobaan berikutnya.

## IV.3 Implementasi Data dan Struktur Berkas

Bagian ini menguraikan realisasi fisik dari perancangan sistem yang mencakup spesifikasi struktur direktori proyek, implementasi modul perangkat lunak, serta skema data atau *data schema* yang digunakan. Implementasi ini dirancang untuk menjamin integritas data eksperimen dan mendukung prinsip keterulangan riset atau *reproducibility*, di mana seluruh artefak data diorganisasikan secara sistematis untuk memfasilitasi audit dan analisis lanjutan.

### IV.3.1 Organisasi Direktori Proyek

Implementasi sistem diorganisasikan dalam struktur direktori hierarkis yang memisahkan kode sumber, konfigurasi, data mentah, dan hasil keluaran guna menjaga modularitas sistem. Struktur direktori proyek didefinisikan sebagai berikut:

1. *Direktori akar.*
   Memuat skrip orkestrator utama dan utilitas eksekusi lainnya yang menjadi titik masuk aplikasi. Direktori ini berfungsi sebagai lapisan kontrol tempat pengguna memulai jalannya eksperimen.
2. *Direktori config.*
   Direktori ini menyimpan seluruh konfigurasi teknis sistem, termasuk pengaturan kredensial API (*Application Programming Interface*) untuk penyedia model seperti Moonshot AI atau OpenRouter. Pemisahan konfigurasi sensitif dari kode sumber utama dilakukan untuk menjaga keamanan dan memudahkan penyesuaian parameter lingkungan tanpa mengubah logika program.
3. *Direktori inputs.*
   Berfungsi sebagai penyimpanan sentral untuk seluruh aset data statis yang diperlukan sebelum eksperimen dijalankan. Di dalamnya terdapat sub-direktori atau berkas untuk definisi *persona* serta *dataset benchmark* standar seperti GSM8K dan MMLU-Redux yang telah divalidasi formatnya.
4. *Direktori results.*
   Direktori ini merupakan pusat penyimpanan seluruh artefak keluaran eksperimen. Di dalamnya terdapat sub-direktori `logs` yang menyimpan hasil eksekusi atau *runtime logs* per sesi secara granular dalam format JSON. Selain itu, direktori ini juga menyimpan hasil analisis teragregasi dan laporan akhir dalam format tabel yang dihasilkan dari pemrosesan data log tersebut.
5. *Direktori src.*
   Memuat seluruh kode sumber perangkat lunak (*source code*) yang ditulis dalam bahasa Python. Di dalamnya terdapat modul-modul fungsional seperti orkestrator eksekusi, klien API, pemantau status, dan mesin analisis data.

### IV.3.2 Implementasi Modul Perangkat Lunak

Logika sistem diimplementasikan ke dalam serangkaian skrip yang diklasifikasikan menjadi empat subsistem fungsional utama untuk memisahkan tanggung jawab pemrosesan.

1. *Subsistem orkestrasi eksekusi.*
   Subsistem ini berfungsi sebagai mesin utama yang menggerakkan alur ekspe-

rimen. Modul orkestrator bertugas memuat konfigurasi dari direktori *config* dan data dari *inputs*, membentuk antrean tugas, dan mendistribusikan beban kerja ke unit pemrosesan. Selain itu, terdapat modul eksekusi spesifik yang menangani inisialisasi parameter untuk penyedia model tertentu.

2. *Subsistem komunikasi antarmuka.*
   Interaksi dengan model bahasa ditangani oleh modul pembungkus klien atau *client wrapper*. Modul ini mengenkapsulasi kompleksitas komunikasi jaringan, termasuk pembentukan pesan JSON, otentikasi menggunakan kunci dari direktori *config*, dan penanganan respons. Sebelum eksperimen dimulai, modul validasi koneksi dijalankan untuk memverifikasi validitas kredensial dan aksesibilitas *endpoint*.
3. *Subsistem pemantauan dan utilitas.*
   Subsistem ini menjamin stabilitas proses melalui mekanisme pemulihan bencana. Modul pemantau status atau *checkpoint monitor* menyimpan kemajuan eksperimen secara berkala, memungkinkan pemulihan proses dari titik terakhir jika terjadi interupsi. Selain itu, modul pelaporan kemajuan menyediakan visibilitas terhadap status penyelesaian tugas asinkron.
4. *Subsistem analisis data.*
   Setelah data terkumpul di dalam direktori *results/logs*, modul analisis melakukan evaluasi komprehensif terhadap log hasil eksperimen. Modul ini dilengkapi dengan logika *parsing* jawaban kompleks untuk mengekstrak nilai numerik atau pilihan ganda dari respons model, serta melakukan agregasi metrik multidimensi. Hasil evaluasi kemudian ditransformasi oleh modul generator laporan menjadi format tabular standar di direktori *results*.

### IV.3.3 Spesifikasi Artefak Data

Integritas eksperimen dijaga melalui standarisasi format data masukan dan keluaran. Spesifikasi data dibagi menjadi dua kategori entitas utama.

Pertama, *spesifikasi data masukan*. Sistem menerima definisi *persona* dalam format JSON yang memuat atribut pengenal unik dan teks instruksi sistem. Data tugas penalaran juga distandarisasi dalam format JSON yang memuat pasangan atribut pertanyaan dan jawaban referensi. Modul pengunduh data secara otomatis menormalisasi format dataset asli dari sumber eksternal menjadi skema yang kompatibel dengan sistem ini.

Kedua, *spesifikasi data keluaran*. Setiap interaksi model direkam dalam berkas log JSON granular yang disimpan di sub-direktori *results/logs*. Skema log ini diran-

cang untuk menangkap telemetri lengkap yang mencakup empat komponen informasi utama: metadata eksekusi yang berisi parameter model, audit input yang menyimpan salinan *prompt* lengkap, respons model berupa teks jawaban mentah, dan statistik penggunaan yang merinci jumlah token dan latensi waktu. Ketersediaan data granular ini memungkinkan analisis mendalam mengenai dampak beban komputasi dari adopsi *persona* secara presisi.

### IV.3.4 Ilustrasi Berkas Data Eksperimen

Untuk memberikan gambaran konkret mengenai implementasi data yang dibahas sebelumnya, berikut disajikan contoh nyata dari berkas masukan dan keluaran yang digunakan dalam sistem.

#### IV.3.4.1 Contoh Konfigurasi Persona

Berkas `persona_echo.json` pada Kode IV.3 merepresentasikan struktur definisi *persona* implisit yang digunakan sebagai masukan. Data ini memuat atribut sumber asal, identitas numerik, dan teks narasi yang mengandung nuansa gaya bahasa pengguna. Sementara itu, Kode IV.4 menunjukkan bagaimana definisi tersebut ditransformasi menjadi instruksi sistem (*system prompt*) yang siap diinjeksikan ke model.

```
{
  "implicit_persona": {
    "source_file": "inputs/implicits_woman_promt.json",
    "id": 1,
    "text": "Lately I've been feeling a strange mix of emotional exhaustion...
             I've been adjusting my skincare routine over and over...
             It's frustrating how something so small can affect my confidence.
  }
}
```

Gambar IV.3 Definisi Persona Implisit pada `persona_echo.json`

#### IV.3.4.2 Contoh Log Keluaran GSM8K

Berkas `gsm8k_00001.json` (Kode IV.5) dan `gsm8k_00003.json` (Kode IV.6) menunjukkan hasil eksekusi tugas penalaran matematika. Log ini merekam metadata model (`model_id`), latensi (`latency_sec`), serta jejak penalaran (*reasoning trace*) yang dihasilkan oleh model.

Perbedaan struktur log pada Kode IV.6 menunjukkan kemampuan sistem untuk menangkap atribut tambahan seperti `reasoning` (teks pemikiran internal) yang terse-

```
{
  "system_prompt": "The user implicitly expresses the following context and co
                    Lately I've been feeling a strange mix of emotional exhaus
                    [...konteks dilanjutkan...]
                    Before I get back to dealing with it, could you help me..."
  "user_prompt": "I appreciate you listening. Before we start, please acknowle
  "response": {
    "choices": [
      {
        "message": {
          "role": "assistant",
          "content": "I hear you clearly. Dealing with persistent skin issues.
        }
      }
    ]
  }
}
```


Gambar IV.4 Struktur Injeksi Konteks pada `persona_warmup.json`


```
{
  "run": {
    "model_id": "openrouter/bert-nebulon-alpha",
    "question_id": "gsm8k_00001",
    "latency_sec": 5.935
  },
  "input": {
    "question": "Janet's ducks lay 16 eggs per day... How much...",
    "gold_answer": "Janet sells 16 - 3 - 4 = 9... #### 18"
  },
  "response": {
    "choices": [
      {
        "message": {
          "content": "Let's break down the problem step by step...
                     1. Total eggs laid per day: 16...
                     Final answer: 18"
        }
      }
    ],
    "usage": {
      "prompt_tokens": 211,
      "completion_tokens": 197
    }
  }
}
```


Gambar IV.5 Contoh Log Eksekusi `gsm8k_00001.json`

```
{
  "run": {
    "model_id": "nvidia/nemotron-nano-12b-v2-vl:free",
    "question_id": "gsm8k_00003",
    "latency_sec": 15.833
  },
  "response": {
    "choices": [
      {
        "message": {
          "content": "Josh's total cost is $80,000 + $50,000 = $130,000...
                     70000",
          "reasoning": "Okay, let's see. Josh bought a house for $80,000..."
        }
      }
    ],
    "usage": {
      "prompt_tokens": 202,
      "completion_tokens": 867
    }
  }
}
```


Gambar IV.6 Contoh Log Eksekusi `gsm8k_00003.json` dengan *Reasoning Trace*

dia pada model-model tertentu seperti Nvidia Nemotron, yang krusial untuk analisis transparansi penalaran.

## IV.4 Rancangan Evaluasi dan Metrik

Tahap akhir dari desain solusi adalah penetapan mekanisme evaluasi untuk mengukur dampak variasi *user persona* terhadap perilaku model. Rancangan ini mendefinisikan metrik kuantitatif yang digunakan untuk menilai performansi penalaran dan efisiensi komputasi, serta spesifikasi format data analisis yang dihasilkan untuk keperluan uji statistik.

### IV.4.1 Metrik Performansi Penalaran

Evaluasi kualitas jawaban model didasarkan pada ketepatan hasil akhir atau *accuracy* terhadap kunci jawaban yang tersedia dalam *dataset*. Mengingat format keluaran model bahasa yang bersifat generatif dan tidak terstruktur, mekanisme evaluasi menerapkan logika pencocokan pola atau *pattern matching* yang ketat.

1. *Akurasi jawaban numerik.*
   Untuk tugas penalaran matematika seperti pada *dataset* GSM8K, metrik utama yang digunakan adalah *Exact Match* pada nilai numerik akhir. Sistem analisis mengekstrak angka terakhir yang dihasilkan model setelah penanda khusus, kemudian membandingkannya dengan nilai kunci jawaban. Jika nilai tersebut identik secara matematis, maka respons dianggap benar (bernilai 1), sebaliknya dianggap salah (bernilai 0).
2. *Akurasi jawaban pilihan ganda.*
   Untuk tugas pengetahuan umum seperti pada MMLU-Redux, evaluasi dilakukan dengan mendeteksi pemilihan opsi jawaban (A, B, C, atau D). Sistem memvalidasi apakah model secara eksplisit memilih opsi yang sesuai dengan kebenaran dasar atau *ground truth*. Akurasi dihitung sebagai persentase jawaban benar dari total pertanyaan yang diajukan untuk setiap kombinasi model dan *persona*.

### IV.4.2 Metrik Efisiensi Komputasi

Selain akurasi, penelitian ini juga mengevaluasi dampak *persona* terhadap beban komputasi model. Indikator efisiensi diukur melalui dua parameter telemetri utama yang direkam selama eksperimen berlangsung.

1. *Verbositas dan penggunaan token ternormalisasi.*
   Metrik ini mengukur jumlah token yang dihasilkan model dalam menjawab

sebuah pertanyaan tugas atau *completion tokens*. Untuk menjamin validitas perbandingan antar-*persona*, dilakukan mekanisme normalisasi dalam perhitungan token. Token yang dialokasikan untuk fase inisialisasi atau *warm-up* serta token *echo* dieksklusi secara total. Pengukuran hanya difokuskan pada token yang dibangkitkan untuk menjawab soal *benchmark*. Pendekatan ini memastikan bahwa metrik efisiensi secara murni merefleksikan biaya kognitif model dalam menyelesaikan masalah.

2. *Latensi inferensi tugas.*
   Waktu yang dibutuhkan model untuk menghasilkan respons penuh diukur dalam satuan detik. Serupa dengan perhitungan token, latensi yang diukur adalah durasi waktu eksekusi spesifik untuk menjawab pertanyaan *benchmark*. Peningkatan latensi pada *persona* tertentu dapat mengindikasikan bahwa model memerlukan upaya komputasi yang lebih tinggi untuk menyelaraskan respons dengan batasan peran yang diberikan.

### IV.4.3 Format Data Analisis

Untuk memfasilitasi analisis komparatif yang komprehensif, data log mentah ditransformasi menjadi format tabular terstruktur. Berdasarkan implementasi sistem, struktur data hasil dibagi menjadi dua tingkat granularitas.

1. *Data hasil granular.*
   Berkas ini menyimpan rekam jejak setiap butir soal secara mendetail sebagaimana terlihat pada berkas *grok_4_1_results.csv*. Atribut kolom mencakup identitas pertanyaan (*Question ID*), *persona* yang digunakan, status kebenaran jawaban (*Correct*), serta telemetri per pertanyaan yang meliputi latensi (*Latency*) dan jumlah token jawaban (*Completion Tokens*).
2. *Data hasil teragregasi.*
   Berkas ringkasan seperti *summary_all_models.csv* digunakan untuk perbandingan tingkat tinggi. Atribut kolom mencakup dimensi eksperimen yaitu nama model dan tipe *persona* (eksplisit/implisit), serta metrik rata-rata yang terdiri dari *Accuracy*, *Average Latency*, dan *Average Token Usage*.

### IV.4.4 Ilustrasi Data Hasil Eksperimen

Untuk memberikan gambaran konkret mengenai bentuk data yang dihasilkan oleh sistem evaluasi, Tabel IV.2 menyajikan sampel data hasil granular yang diekstraksi dari hasil pengujian model Grok 4.1 pada tugas GSM8K. Tabel ini memperlihatkan bagaimana variasi *persona* memengaruhi latensi dan penggunaan token pada soal

yang berbeda.

Tabel IV.2 Sampel Data Hasil Granular (Grok 4.1)

| *Question ID* | *Persona* | *Correct* | *Latency (s)* | *C. Tokens* |
|---|---|---|---|---|
| gsm8k_00001 | Neutral | TRUE | 4.21 | 180 |
| gsm8k_00001 | Explicit Man | TRUE | 4.50 | 195 |
| gsm8k_00002 | Neutral | FALSE | 3.80 | 140 |
| gsm8k_00002 | Explicit Man | TRUE | 5.10 | 210 |

*Keterangan: C. Tokens merujuk pada Completion Tokens.*

Sementara itu, Tabel IV.3 menampilkan format data teragregasi yang digunakan untuk analisis komparatif antar-model sebagaimana terdapat pada berkas *summary_all_models.csv*.

Tabel IV.3 Sampel Data Hasil Teragregasi (Ringkasan)

| *Model* | *Persona* | *Accuracy* | *Avg Latency* | *Avg Tokens* |
|---|---|---|---|---|
| Grok 4.1 | Neutral | 0.88 | 4.5s | 180 |
| Grok 4.1 | Explicit Man | 0.89 | 4.6s | 175 |
| Bert Nebulon | Neutral | 0.72 | 3.2s | 140 |
| Bert Nebulon | Explicit Man | 0.70 | 3.5s | 155 |

# 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.* https://huggingface.co/datasets/edinburgh-dawg/mmlu-redux-2.0. Subset MMLU yang dianotasi ulang secara manual untuk 57 subjek, 100 soal per subjek.

Gema, Aryo Pradipta, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, 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". *Proceedings of the 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.