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 4 Desember 2025

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

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE

BAB I

PENDAHULUAN

I.1 Latar Belakang

Kemajuan dalam pengembangan *large language model* (LLM) dalam beberapa tahun terakhir telah mengubah cara sistem komputasi memahami, memproses, dan menghasilkan bahasa alami. Model seperti GPT, LLaMA, Grok, dan Gemini dilatih menggunakan korpus berskala besar dan mampu menyelesaikan berbagai tugas mulai dari penalaran numerik hingga interpretasi skenario sosial (**jurafsky2023slp3**). Pada sejumlah benchmark terstandardisasi, model-model tersebut dapat memberikan jawaban yang akurat dan relevan. Namun, peningkatan kemampuan ini belum sepenuhnya diikuti oleh konsistensi perilaku model dalam percakapan. Perubahan kecil dalam cara pertanyaan disampaikan sering kali menghasilkan respons yang berbeda, meskipun tugas yang diberikan tetap sama (Zhou dkk. 2023).

Fenomena lain yang semakin banyak dibahas dalam penelitian mutakhir adalah bahwa perilaku model tidak hanya dipengaruhi oleh isi instruksi, tetapi juga oleh cara model memersepsi identitas pengguna. Studi mengenai bias penalaran implisit menunjukkan bahwa deskripsi singkat mengenai pengguna dapat mengubah pola penalaran model, termasuk pada tugas-tugas yang tidak memiliki muatan sosial, seperti penalaran numerik atau penyelesaian masalah dasar (Gupta dkk. 2024). Perubahan tersebut mencakup variasi langkah penyelesaian, tingkat kehati-hatian, ataupun kecenderungan preferensi tertentu terhadap kelompok sosial.

Selain persona yang dinyatakan secara langsung, beberapa penelitian menemukan bahwa model dapat mengasosiasikan isyarat linguistik halus—seperti pilihan kata, tingkat formalitas, atau gaya pertanyaan—dengan karakteristik tertentu dari pengguna (Tseng dkk. 2024). Asosiasi ini kemudian berpotensi memengaruhi strategi penyelesaian yang dipilih model, termasuk variasi pada langkah-langkah penalaran yang biasanya tercermin dalam *chain-of-thought*.

Penelitian dalam pemodelan pengguna juga menunjukkan bahwa identitas pengguna—meliputi usia, latar belakang profesional, maupun pengalaman tertentu—dapat memberikan pengaruh terhadap pola respons model (Naous, Roziere, dkk. 2025). Dalam penelitian ini, identitas pengguna direpresentasikan melalui persona yang dibentuk secara eksplisit maupun implisit di dalam prompt. Pendekatan tersebut digunakan untuk mengkaji bagaimana model membangun asumsi mengenai pengguna dan bagaimana asumsi tersebut tercermin pada keluaran model dalam berbagai skenario tugas.

Meskipun terdapat sejumlah temuan penting, penelitian terdahulu masih memiliki keterbatasan. Sebagian besar hanya melibatkan jumlah model yang terbatas, ruang persona yang sempit, atau cakupan tugas yang relatif kecil. Belum banyak penelitian yang secara sistematis membandingkan persona eksplisit dan implisit pada berbagai model dan berbagai jenis penalaran dalam kerangka eksperimen yang konsisten. Selain itu, penelitian mengenai perbedaan antara pendekatan persona berbasis pengguna ("your user is...") dan pendekatan berbasis model ("you are...") juga masih terbatas, padahal kedua bentuk framing tersebut berpotensi menghasilkan respons yang berbeda. Dalam konteks ini, studi seperti HELM (Liang, Bommasani, dkk. 2023) menegaskan bahwa model sensitif terhadap variasi konteks yang tampak kecil, sehingga evaluasi terstruktur menjadi semakin penting.

Keterbatasan tersebut semakin relevan mengingat penerapan model bahasa pada berbagai bidang yang sensitif terhadap identitas pengguna, seperti pendidikan, layanan kesehatan, dan sistem rekomendasi. Ketidakstabilan respons yang dipicu oleh variasi cara model memersepsi pengguna dapat mengurangi keandalan sistem dan menimbulkan bias. Selain itu, variasi hasil antar-*run* pada tugas yang sama menunjukkan perlunya mekanisme evaluasi yang terstruktur dan dapat direproduksi (Turpin dkk. 2023; Cobbe dkk. 2021).

Berangkat dari kebutuhan tersebut, penelitian ini disusun untuk mengevaluasi pengaruh persona eksplisit dan implisit melalui eksperimen terstruktur pada berbagai model dan jenis tugas penalaran. Penelitian ini memanfaatkan pendekatan *spec-driven experiment orchestration* yang memungkinkan pelaksanaan kombinasi persona–model–benchmark secara konsisten dan dapat diulang. Dengan pendekatan tersebut, penelitian ini diharapkan dapat memberikan gambaran yang lebih jelas mengenai bagaimana model menafsirkan identitas pengguna dan bagaimana penafsiran tersebut memengaruhi jawaban dalam berbagai konteks tugas.

I.2 Rumusan Masalah

Penelitian sebelumnya menunjukkan bahwa persona, baik yang diberikan secara eksplisit maupun yang tersirat dari gaya bahasa, dapat memengaruhi cara model menyusun penalaran dan menghasilkan jawaban (Gupta dkk. 2024; Tseng dkk. 2024; Naous, Roziere, dkk. 2025). Namun, kajian yang ada masih terbatas pada jumlah model yang sedikit, ragam persona yang sempit, serta jenis tugas yang belum cukup mencerminkan variasi penalaran yang lebih luas. Kondisi ini menunjukkan perlunya evaluasi yang lebih menyeluruh untuk memahami bagaimana persona memengaruhi perilaku model dalam konteks multi-tugas dan multi-model.

Berdasarkan uraian pada bagian sebelumnya, penelitian ini merumuskan beberapa pertanyaan utama sebagai berikut.

1. Sejauh mana persona yang diberikan secara eksplisit maupun yang muncul secara implisit memengaruhi proses penalaran model pada berbagai jenis tugas, khususnya penalaran numerik dan tugas multi-topik?
2. Bagaimana variasi persona tersebut membentuk karakter keluaran model dan memunculkan pola bias tertentu, termasuk bias sosial maupun preferensi jawaban?
3. Bagaimana perbedaan respons antar model dapat menggambarkan tingkat sensitivitas dan ketahanan masing-masing model terhadap variasi persona dalam suatu kerangka evaluasi yang disusun secara terstruktur?

I.3 Tujuan Penelitian

Tujuan penelitian ini disusun sebagai tindak lanjut dari rumusan masalah yang telah dijelaskan sebelumnya. Secara umum, penelitian ini bertujuan untuk memperoleh pemahaman yang lebih jelas mengenai pengaruh persona terhadap perilaku dan penalaran *large language model*. Secara khusus, penelitian ini bertujuan untuk:

1. Menganalisis sejauh mana persona yang diberikan secara eksplisit maupun yang muncul secara implisit memengaruhi proses *reasoning* model pada berbagai jenis tugas.
2. Mengidentifikasi perubahan karakter keluaran dan pola *bias* yang muncul pada model sebagai akibat dari variasi persona.
3. Mengevaluasi perbedaan respons antar model untuk menilai tingkat *sensitivity* dan *robustness* masing-masing model terhadap variasi persona dalam pengaturan eksperimen yang disusun secara terstruktur.

I.4 Batasan Masalah

Batasan masalah diperlukan agar ruang lingkup penelitian tetap jelas dan terarah. Penelitian ini tidak mencakup seluruh aspek perilaku *large language model*, tetapi memfokuskan kajian pada bagaimana variasi persona memengaruhi respons model pada sejumlah tugas penalaran. Adapun batasan penelitian ini adalah sebagai berikut.

1. Penelitian hanya mempertimbangkan dua bentuk persona yang berorientasi pada pengguna, yaitu persona eksplisit yang dinyatakan secara langsung di dalam prompt, serta persona implisit yang muncul dari variasi gaya bahasa dan cara pengguna menyampaikan pertanyaan. Kajian ini tidak mencakup *role-playing persona* yang menetapkan identitas tertentu pada model, maupun pendekatan *personalization* yang bergantung pada riwayat atau profil pengguna.
2. Model yang ditelaah terbatas pada model bahasa berbasis teks yang tersedia melalui antarmuka API. Model multimodal serta model yang memerlukan proses *fine-tuning* atau pelatihan ulang tidak menjadi bagian dari penelitian ini.
3. Evaluasi dilakukan pada tugas-tugas berbasis teks, meliputi penalaran numerik, penalaran logis, pertanyaan pengetahuan umum, serta skenario sosial dan moral. Penelitian ini tidak membahas tugas multimodal maupun tugas berbasis *speech*.
4. Penilaian terhadap respons model dilakukan melalui evaluasi otomatis dan analisis komparatif. Penelitian tidak melibatkan penilaian dengan partisipan manusia.
5. Seluruh eksperimen dijalankan melalui pendekatan *prompt-based evaluation* tanpa melakukan perubahan terhadap parameter internal model.
6. Analisis bias dibatasi pada *human bias* yang timbul sebagai konsekuensi variasi persona. Penelitian tidak mengevaluasi bias yang berasal dari data pelatihan model atau faktor struktur model lainnya.

I.5 Metodologi Penelitian

Penelitian ini menggunakan pendekatan eksperimental berbasis pemanggilan model melalui prompt untuk melihat bagaimana persona memengaruhi respons sejumlah *large language model*. Metodologi dirancang agar alur evaluasi jelas dan dapat dijalankan kembali apabila diperlukan. Tahapan penelitian disajikan sebagai berikut.

1. Perumusan spesifikasi eksperimen.

Tahap ini diawali dengan menyusun dokumen spesifikasi yang memetakan kombinasi persona, model, bentuk interaksi, dan jenis tugas yang akan diuji. Spesifikasi tersebut dipakai sebagai acuan sehingga pelaksanaan eksperimen berjalan dengan alur yang tetap.

2. Penyusunan persona eksplisit dan implisit.
   Persona eksplisit dituliskan secara langsung di dalam prompt, sedangkan persona implisit dibangun melalui variasi gaya bahasa pengguna tanpa menyebutkan identitas secara eksplisit. Kedua bentuk persona digunakan untuk melihat bagaimana model memahami karakter pengguna dari konteks yang berbeda.
3. Pemilihan model dan ruang evaluasi.
   Penelitian menggunakan beberapa model bahasa berbasis teks yang tersedia melalui API tanpa proses *fine-tuning*. Tugas yang digunakan mencakup penalaran numerik, penalaran logis, pertanyaan pengetahuan umum, serta skenario sosial dan moral.
4. Pelaksanaan eksperimen terotomatisasi.
   Setiap kombinasi persona, model, dan tugas dieksekusi menggunakan pendekatan *prompt-based evaluation*. Seluruh proses dijalankan secara otomatis untuk mengurangi variasi yang tidak diperlukan dan menjaga alur pengujian tetap seragam.
5. Pengolahan respons dan analisis perbandingan.
   Respons model dicatat dan dianalisis berdasarkan ketepatan jawaban serta pola perubahan respons yang muncul akibat perbedaan persona. Perbandingan antar model dilakukan untuk melihat sejauh mana masing-masing model peka terhadap perubahan persona.
6. Analisis bias.
   Analisis difokuskan pada *human bias* yang muncul selama proses tanya jawab akibat variasi persona. Penelitian ini tidak meninjau bias yang berasal dari data pelatihan atau arsitektur model.

Metodologi ini menjadi dasar untuk pelaksanaan eksperimen dan pembahasan pada bab selanjutnya.

BAB II

STUDI LITERATUR

II.1 Large Language Model

Large language model (LLM) adalah model berbasis arsitektur transformator yang dilatih menggunakan korpus teks dalam jumlah sangat besar. Melalui proses pelatihan ini, model mempelajari hubungan antar-token, pola semantik, serta variasi penggunaan bahasa yang umum ditemukan pada teks manusia (Bommasani, Hudson, Adeli, dkk. 2021). Dengan kemampuan tersebut, LLM dapat menghasilkan jawaban yang relevan meskipun hanya diberikan instruksi berbasis teks.

Salah satu ciri penting LLM adalah sensitivitasnya terhadap konteks. Model tidak hanya melihat makna literal suatu kata, tetapi juga memperhatikan gaya penulisan, struktur kalimat, dan isyarat pragmatik yang terdapat pada instruksi. Penelitian menunjukkan bahwa perubahan kecil dalam cara instruksi ditulis—misalnya perbedaan nada atau tingkat formalitas—dapat menghasilkan jawaban yang berbeda meskipun maksud pengguna tetap sama (Zhou dkk. 2023).

Sifat ini juga berpengaruh pada bentuk penalaran yang dihasilkan. Turpin (Turpin dkk. 2023) menemukan bahwa langkah penalaran LLM dapat berubah hanya karena variasi kecil dalam formulasi instruksi. Temuan ini menunjukkan bahwa LLM tidak selalu mengikuti jalur penalaran yang konsisten, tetapi menyesuaikannya dengan konteks linguistik yang diterima saat inferensi dilakukan.

Selain peka terhadap bahasa, LLM juga mempelajari pola sosial dari data pelatihan. Weidinger (Weidinger dkk. 2021) menunjukkan bahwa model dapat menyerap preferensi sosial atau bias yang muncul dalam korpus pelatihan. Dalam praktiknya, gaya penulisan pengguna dapat ditafsirkan sebagai sinyal identitas tertentu. Studi mengenai *role-playing* juga menunjukkan bahwa isyarat persona dapat memengaruhi gaya penjelasan, tingkat kehati-hatian, dan pola penalaran yang dihasilkan model

(Tseng dkk. 2024).

Karena persona merupakan variasi konteks linguistik yang dapat mengubah cara model menafsirkan instruksi, pemahaman mengenai sensitivitas ini menjadi dasar penting bagi penelitian. Pengaruh persona terhadap penalaran model perlu dianalisis secara sistematis pada berbagai LLM dengan ukuran dan karakteristik yang berbeda.

II.2 Persona dalam Interaksi Model Bahasa

Persona dalam konteks LLM merujuk pada karakteristik identitas yang tercermin melalui cara seseorang menulis atau berkomunikasi. Ciri ini dapat berupa pilihan kosakata, tingkat formalitas, struktur kalimat, atau pola penyampaian informasi. Dalam komunikasi manusia, perbedaan persona membantu lawan bicara menafsirkan maksud dan menyesuaikan respons. Hal serupa juga terjadi pada LLM karena model belajar dari data yang memuat berbagai gaya komunikasi (Tseng dkk. 2024).

Dalam interaksi dengan LLM, persona dapat muncul dalam dua bentuk, yaitu eksplisit dan implisit. Persona eksplisit muncul ketika identitas disebutkan secara langsung, seperti "Sebagai mahasiswa teknik informatika…". Informasi seperti ini memberikan sinyal identitas yang jelas sehingga model dapat menyesuaikan gaya atau struktur jawaban. Gupta (Gupta dkk. 2024) menunjukkan bahwa penugasan persona eksplisit dapat menghasilkan bentuk penalaran yang berbeda meskipun tugas yang diberikan sama.

Persona implisit muncul ketika model menyimpulkan identitas pengguna dari gaya penulisan tanpa adanya pernyataan langsung. Misalnya, gaya formal sering diasosiasikan dengan konteks akademis, sedangkan gaya santai lebih banyak ditemukan pada percakapan sehari-hari. Ketika pola tertentu muncul dalam instruksi, model dapat menafsirkannya sebagai identitas tertentu dan menyesuaikan jawabannya. Tseng (Tseng dkk. 2024) menunjukkan bahwa inferensi identitas seperti ini dapat terjadi hanya dari perbedaan pola bahasa.

Persona dapat memengaruhi dua aspek utama dalam respons model. Pertama, bentuk jawaban. Variasi persona dapat mengubah panjang penjelasan, pilihan kosakata, atau tingkat formalitas. Model cenderung menyesuaikan gaya jawaban agar sesuai dengan persona yang muncul pada instruksi.

Kedua, penalaran. Karena LLM peka terhadap formulasi instruksi (Zhou dkk. 2023) dan dapat menghasilkan langkah penalaran yang berbeda meskipun tugasnya sama

(Turpin dkk. 2023), perubahan persona dapat memicu variasi dalam cara model mencapai kesimpulan. Persona tertentu dapat membuat model menyusun penalaran yang lebih panjang atau lebih berhati-hati, sedangkan persona lain dapat menghasilkan penalaran yang lebih ringkas.

Penelitian mengenai bias juga menunjukkan bahwa persona dapat berinteraksi dengan pola sosial yang dipelajari model. Weidinger (Weidinger dkk. 2021) menunjukkan bahwa model dapat memperlakukan persona tertentu secara berbeda jika pola tersebut sering muncul dalam data pelatihan. Dalam beberapa situasi, persona dapat menggeser preferensi model dalam memilih sudut pandang atau jenis penjelasan yang diberikan (Gupta dkk. 2024).

Persona bukan sekadar tambahan informasi dalam instruksi, tetapi merupakan bagian dari konteks yang dipertimbangkan model ketika membentuk jawaban. Karena persona dapat mengubah gaya maupun penalaran model, analisis pengaruh persona menjadi penting untuk memahami konsistensi respons LLM pada kondisi identitas pengguna yang berbeda.

II.3 Pengaruh Persona terhadap Perilaku LLM

Persona dapat memengaruhi cara LLM memahami instruksi dan menghasilkan jawaban. Efek ini muncul karena model belajar dari data pelatihan yang berisi berbagai gaya bahasa dan situasi komunikasi. Ketika instruksi ditulis dengan gaya tertentu, model dapat menafsirkannya sebagai sinyal identitas dan menyesuaikan cara menjawab.

II.3.1 Pengaruh Persona terhadap Penalaran

Penelitian menunjukkan bahwa langkah penalaran LLM tidak selalu konsisten. Gupta (Gupta dkk. 2024) menemukan bahwa menambahkan persona eksplisit ke dalam instruksi dapat membuat model menghasilkan penalaran yang berbeda meskipun tugasnya sama. Perbedaan ini dapat terlihat pada urutan penjelasan, tingkat kehati-hatian, atau cara model menyusun argumen.

Pada persona implisit, perubahan penalaran muncul dari hal-hal yang lebih halus, seperti pilihan kata atau tingkat formalitas. Instruksi dengan gaya formal sering membuat model memberikan penjelasan yang lebih terstruktur. Sebaliknya, gaya penulisan yang santai dapat memicu jawaban yang lebih ringkas. Temuan Turpin (Turpin dkk. 2023) menunjukkan bahwa perubahan kecil pada formulasi instruksi

dapat mengubah langkah penalaran yang dihasilkan model. Kondisi ini membuat persona menjadi salah satu faktor yang dapat memicu perbedaan tersebut.

II.3.2 Pengaruh Persona terhadap Gaya Respons

Selain penalaran, persona juga memengaruhi cara model menyampaikan jawaban. Tseng (Tseng dkk. 2024) menunjukkan bahwa model dapat menyesuaikan gaya bahasa meskipun identitas pengguna tidak disebutkan secara langsung. Efek ini dapat terlihat dari panjang kalimat, tingkat formalitas, atau nada penjelasan.

Jika model mengaitkan persona tertentu dengan konteks profesional, respons yang diberikan cenderung lebih sistematis dan terstruktur. Sebaliknya, pada persona yang diasosiasikan dengan percakapan santai, jawaban yang muncul biasanya lebih singkat dan langsung.

II.3.3 Faktor yang Memengaruhi Efek Persona

Pengaruh persona dapat menjadi lebih kuat ketika framing instruksi konsisten. Selain itu, jenis tugas juga berperan. Pada tugas yang lebih terbuka, seperti skenario sosial, efek persona cenderung lebih terlihat dibandingkan pada tugas yang memiliki jawaban pasti. Ukuran dan kapasitas model juga memengaruhi sejauh mana persona berdampak terhadap respons. Model yang lebih besar umumnya lebih sensitif terhadap variasi gaya bahasa.

II.4 Bias dalam Respons LLM

Bias pada LLM muncul karena model belajar dari data yang mengandung kecenderungan tertentu. Selain informasi faktual, data pelatihan juga memuat pola sosial, stereotip, atau kebiasaan bahasa yang umum digunakan. Pola tersebut dapat terbawa ke dalam jawaban model.

II.4.1 Bentuk-bentuk Bias

Weidinger (Weidinger dkk. 2021) menunjukkan bahwa model dapat meniru stereotip yang ada pada data pelatihan. Bias seperti ini disebut bias representasional, misalnya ketika model menggambarkan suatu profesi atau kelompok sosial dengan cara yang tidak seimbang.

Selain itu, terdapat bias inferensial, yaitu ketika model menarik kesimpulan berdasarkan asosiasi yang tidak relevan. Model dapat menambahkan detail yang tidak

disebutkan pengguna hanya karena pola tersebut sering muncul dalam data pelatihan.

Bias penalaran juga dapat muncul. Gupta (Gupta dkk. 2024) menemukan bahwa persona tertentu dapat mendorong model menggunakan pola penjelasan tertentu yang tidak selalu muncul pada instruksi netral.

II.4.2 Dampak Bias terhadap Keluaran

Bias dapat memengaruhi ketepatan jawaban. Model dapat memberikan respons yang terdengar meyakinkan tetapi tidak sesuai dengan konteks yang diminta. Bias juga dapat memengaruhi panjang atau gaya penjelasan. Dalam beberapa kasus, model memberikan penjelasan lebih rinci kepada persona tertentu dan lebih singkat kepada persona lainnya.

Bias juga dapat memperkuat pola sosial tertentu secara tidak langsung. Misalnya, pemilihan kata atau nada penjelasan dapat mencerminkan kecenderungan tertentu tanpa disadari.

II.4.3 Kaitannya dengan Persona

Persona dapat memperkuat atau mengubah bias tersebut. Pada persona eksplisit, penyebutan identitas dapat memicu asosiasi tertentu yang pernah muncul dalam data pelatihan. Pada persona implisit, perubahan gaya bahasa dapat membuat model menafsirkan identitas tertentu meskipun tidak disebutkan secara langsung (Tseng dkk. 2024).

Penelitian Zhou (Zhou dkk. 2023) dan Turpin (Turpin dkk. 2023) menunjukkan bahwa LLM sangat sensitif terhadap perubahan formulasi instruksi. Karena persona merupakan bagian dari formulasi tersebut, variasi persona dapat menyebabkan perubahan dalam penjelasan atau penalaran yang dihasilkan model.

Pengaruh ini lebih terlihat pada tugas yang bersifat terbuka, seperti skenario sosial. Pada konteks seperti ini, ruang interpretasi yang lebih luas membuat bias dan persona lebih mudah memengaruhi hasil akhir.

II.5 Evaluasi Penalaran dan Benchmark

Evaluasi terhadap LLM umumnya dilakukan menggunakan benchmark yang dirancang untuk mengukur kemampuan penalaran dan pemahaman model secara lebih

terstruktur. Benchmark membantu memberikan gambaran mengenai performa model pada tugas yang bersifat konsisten dan terukur. Dalam penelitian ini, dua benchmark digunakan untuk menilai bagaimana persona dapat memengaruhi cara model menjawab, yaitu GSM8K dan MMLU-Redux.

II.5.1 GSM8K

GSM8K adalah kumpulan soal matematika tingkat sekolah dasar yang dirancang untuk menguji kemampuan penalaran numerik (Cobbe dkk. 2021). Setiap soal biasanya membutuhkan beberapa langkah pemikiran sederhana, seperti memahami konteks, melakukan perhitungan dasar, dan menarik kesimpulan. Bagi manusia, soal-soal ini relatif mudah, tetapi bagi LLM, benchmark ini menantang karena model harus menyusun langkah penyelesaian yang runtut.

Benchmark ini digunakan dalam penelitian untuk melihat apakah variasi persona dapat memengaruhi cara model membentuk langkah penalaran tersebut. Misalnya, persona tertentu dapat membuat model memberikan penjelasan lebih panjang, sementara persona lain mendorong model untuk menjawab lebih singkat. Dengan demikian, GSM8K memberikan konteks yang jelas untuk mengamati perubahan pada pola penalaran model.

II.5.2 MMLU-Redux

MMLU-Redux merupakan versi kurasi ulang dari benchmark MMLU yang berisi pertanyaan dari berbagai bidang pengetahuan (Edinburgh Dataset Analytics Working Group 2024). Tidak seperti GSM8K yang terfokus pada matematika dasar, MMLU-Redux mencakup berbagai kategori seperti sains, humaniora, dan ilmu sosial. Soal-soal dalam benchmark ini menguji kemampuan model dalam memahami konsep dan memilih jawaban yang paling tepat berdasarkan pengetahuan umum.

Benchmark ini digunakan dalam penelitian untuk melihat bagaimana persona dapat memengaruhi pilihan jawaban model, terutama pada pertanyaan yang memerlukan pemahaman konsep dan penalaran tingkat menengah. Karena format soal bersifat pilihan ganda, MMLU-Redux memberikan lingkungan evaluasi yang lebih terkontrol, sehingga perbedaan respons yang muncul lebih mudah diamati dari sisi persona.

II.5.3 Tantangan Evaluasi Berbasis Persona

Penggunaan benchmark dalam penelitian persona memiliki beberapa tantangan. Tantangan pertama adalah memastikan bahwa perubahan jawaban benar-benar dise-

babkan oleh persona, bukan oleh perbedaan formulasi instruksi. Karena LLM peka terhadap gaya penulisan (Zhou dkk. 2023), evaluasi perlu dilakukan dengan struktur prompt yang konsisten.

Tantangan berikutnya adalah variasi hasil yang terjadi antarpemanggilan model. LLM dapat menghasilkan jawaban berbeda meskipun instruksi yang diberikan sama (Turpin dkk. 2023). Oleh karena itu, proses evaluasi dilakukan secara terotomatisasi dan terstandarisasi agar hasil yang diperoleh lebih dapat dibandingkan.

Harapannya, benchmark GSM8K dan MMLU-Redux memberikan dasar yang jelas untuk melihat bagaimana persona dapat memengaruhi penalaran dan pilihan jawaban model dalam dua konteks yang berbeda, yaitu penalaran numerik dan pengetahuan umum.

II.6 Penelitian Terdahulu dan Kesenjangan Penelitian

Pembahasan mengenai persona dan perilaku LLM telah dibahas dalam beberapa penelitian sebelumnya. Secara umum, penelitian-penelitian tersebut menunjukkan bahwa identitas pengguna—baik yang dinyatakan secara langsung maupun tersirat melalui gaya penulisan—dapat memengaruhi cara model menghasilkan jawaban. Namun, sebagian besar studi masih terbatas pada jenis model atau bentuk persona tertentu sehingga gambaran mengenai pengaruh persona secara lebih luas belum banyak diuraikan.

II.6.1 Ringkasan Literatur Terkait

Gupta (Gupta dkk. 2024) menunjukkan bahwa pemberian persona eksplisit dapat mengubah langkah penalaran model pada tugas yang sama. Temuan ini memperlihatkan bahwa model tidak hanya memproses isi instruksi, tetapi juga memperhatikan informasi identitas yang disisipkan ke dalam prompt.

Tseng (Tseng dkk. 2024) menyoroti persona implisit yang muncul dari pilihan kata dan gaya penulisan. Dalam banyak kasus, model menafsirkan pola bahasa tersebut sebagai sinyal identitas tertentu dan menyesuaikan struktur responsnya. Studi ini memperlihatkan bahwa persona dapat terbentuk bahkan tanpa penyebutan identitas secara langsung.

Turpin (Turpin dkk. 2023) menemukan bahwa LLM dapat menghasilkan urutan penalaran yang berbeda hanya karena perubahan kecil dalam formulasi instruksi. Temuan ini menunjukkan bahwa penalaran model sangat dipengaruhi oleh konteks

linguistik yang diterima saat inferensi.

Dalam konteks bias, Weidinger (Weidinger dkk. 2021) menunjukkan bahwa model dapat meniru pola sosial atau stereotip yang ada dalam data pelatihan. Kondisi ini relevan ketika menilai pengaruh persona karena identitas tertentu dapat memperkuat pola bias yang sudah ada.

II.6.2 Keterbatasan Penelitian Sebelumnya

Meskipun penelitian sebelumnya memberikan kontribusi penting, sebagian besar studi masih memiliki beberapa keterbatasan. Pertama, banyak penelitian hanya menguji sedikit model sehingga belum memberikan gambaran mengenai bagaimana pengaruh persona dapat berbeda antar-LLM. Kedua, jumlah persona yang digunakan umumnya terbatas sehingga variasi efek persona belum terobservasi secara lebih luas. Ketiga, sebagian penelitian hanya menguji sedikit jenis tugas, padahal persona dapat memengaruhi model secara berbeda pada tugas numerik, pengetahuan umum, atau skenario sosial. Selain itu, tidak semua penelitian menggunakan kerangka evaluasi yang terstandarisasi sehingga sulit memastikan bahwa perubahan jawaban benar-benar disebabkan oleh persona.

II.6.3 Posisi dan Kontribusi Penelitian Ini

Penelitian ini disusun untuk mengatasi keterbatasan tersebut. Berbeda dari sebagian studi sebelumnya, penelitian ini menggunakan beberapa model dan beberapa persona untuk melihat bagaimana keduanya memengaruhi penalaran dan respons model. Penelitian ini juga menggunakan dua benchmark yang berbeda—GSM8K dan MMLU-Redux—agar pengaruh persona dapat diamati pada tugas numerik dan pengetahuan umum.

Selain itu, penelitian ini menggunakan *pipeline* evaluasi yang terotomatisasi sehingga setiap model menerima instruksi yang konsisten. Pendekatan ini membantu memastikan bahwa perbedaan yang muncul benar-benar berasal dari persona dan bukan dari variasi struktur prompt.

Dengan demikian, penelitian ini diharapkan dapat memberikan gambaran yang lebih menyeluruh mengenai bagaimana persona memengaruhi perilaku model bahasa, terutama ketika melibatkan beberapa model dan kategori tugas yang berbeda.

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 perubahan 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 dihadapi 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 eksplisit, sehingga skenario eksperimen dapat direplikasi. KF-02 menekankan penting-

nya 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 ekse-

kusi lintas model dan lintas tugas, dukungan pencatatan hasil dan metadata untuk analisis kuantitatif dan kualitatif, keterulangan (*reproducibility*) proses eksperimen, serta tingkat kerumitan implementasi dan pemeliharaan. Ringkasan perbandingan alternatif ditunjukkan pada Tabel III.4, dengan skala kualitatif rendah, sedang, dan tinggi.

Tabel III.4 Perbandingan alternatif solusi

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

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

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

Pendekatan skrip eksperimen semi terotomatisasi berbasis konfigurasi memberikan keseimbangan yang lebih sesuai. Representasi model, persona, dan tugas dapat diatur dalam direktori konfigurasi yang terpisah dari kode, sementara skrip eksekusi dan analisis ditempatkan dalam direktori tersendiri. Keluaran eksperimen disimpan sebagai berkas JSON pada direktori log dan diolah lebih lanjut menjadi berkas CSV pada direktori hasil. Struktur ini mendukung konsistensi eksekusi, keterulangan eksperimen, dan analisis terukur tanpa memerlukan pembangunan kerangka evaluasi

yang terlalu umum.

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

BAB IV

DESAIN KONSEP SOLUSI

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

IV.1 Desain Konseptual Eksperimen

Bagian ini memaparkan dasar konseptual dari eksperimen yang dikembangkan untuk mengukur pengaruh *user persona* terhadap perilaku model bahasa. Perancangan ini mencakup evaluasi terhadap keterbatasan pendekatan konvensional, prinsip desain sistem terotomatisasi yang diusulkan, serta integrasi persona, model, dan *benchmark* penalaran yang digunakan. Dengan adanya desain konseptual ini, alur eksperimen yang dibahas pada subbab berikutnya menjadi lebih terarah, terukur, dan dapat direplikasi.

IV.1.1 Keterbatasan Model Operasional Konvensional

Pendekatan manual yang lazim digunakan dalam penelitian persona umumnya bergantung pada penulisan instruksi langsung melalui antarmuka percakapan. Meskipun sederhana, pendekatan ini memiliki dua kelemahan metodologis utama yang mengurangi validitas internal penelitian.

Pertama, terjadi instabilitas masukan. Model bahasa sangat sensitif terhadap perubahan kecil pada struktur *prompt*—seperti variasi tanda baca atau perubahan gaya

kalimat—yang dapat menyebabkan keluaran berbeda secara signifikan. Turpin et al. menunjukkan bahwa variasi kecil dalam *framing* dapat menghasilkan rantai penalaran yang tidak konsisten (Turpin dkk. 2023). Ketergantungan pada input manual membuat variasi ini tidak dapat dikendalikan.

Kedua, pendekatan manual tidak menyediakan granularitas data yang memadai. Respons model biasanya hanya dicatat dalam bentuk teks akhir tanpa metadata komputasional seperti latensi atau jumlah token. Padahal, indikator tersebut penting untuk memahami beban kognitif model serta pola penalaran yang muncul pada kondisi persona tertentu (Naous, Roziere, dkk. 2025).

IV.1.2 Sistem Eksperimen Terotomatisasi

Untuk mengatasi keterbatasan tersebut, penelitian ini mengusulkan sistem eksperimen terotomatisasi dengan tiga prinsip desain inti: determinisme masukan, telemetri komprehensif, dan skalabilitas eksekusi.

Pertama, determinisme masukan dicapai dengan menyimpan seluruh persona dalam berkas konfigurasi statis. Setiap *prompt* dibentuk melalui mekanisme injeksi otomatis sehingga stimulus yang diterima model identik hingga tingkat karakter.

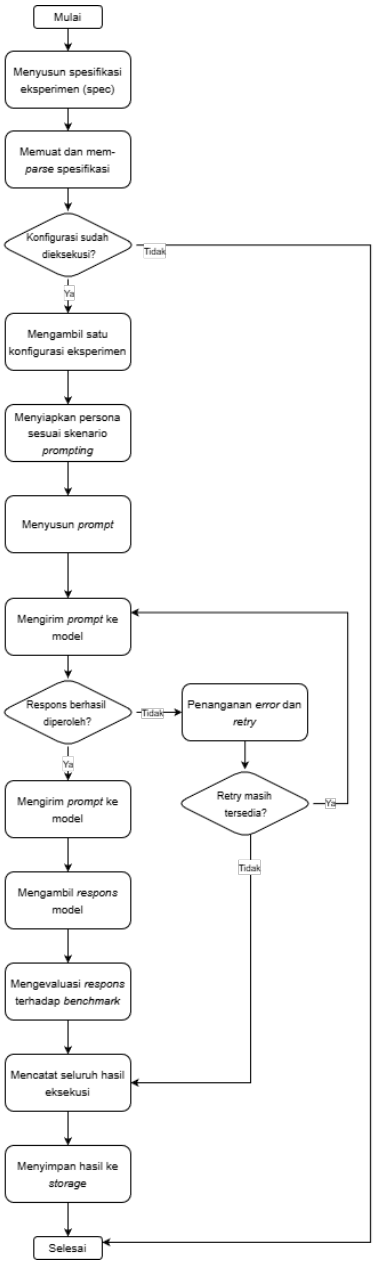
Kedua, sistem mencatat keluaran model secara lengkap dalam format terstruktur (JSON dan CSV). Telemetri yang direkam mencakup teks jawaban, *reasoning trace* (bila tersedia), jumlah token, dan latensi inferensi. Dengan demikian, analisis dapat menilai tidak hanya kebenaran jawaban, tetapi juga pola beban komputasi yang terkait dengan masing-masing persona.

Ketiga, sistem mendukung eksekusi *asynchronous* sehingga ribuan permintaan dapat diproses secara paralel tanpa melampaui batas layanan. Pendekatan ini meningkatkan cakupan eksperimen sekaligus mengurangi waktu eksekusi total.

Model konseptual dari sistem usulan ditampilkan pada Gambar IV.1, yang dalam penomoran dokumen dicantumkan sebagai **Gambar IV.2 Model Konseptual Sistem Eksperimen Terotomatisasi**.

IV.1.3 Analisis Komparatif Metodologis

Peralihan dari pendekatan *Existing* ke *Proposed* bukan sekadar peningkatan efisiensi, melainkan juga penguatan validitas ilmiah. Tabel IV.1 merangkum perbedaan metodologis utama antara kedua pendekatan tersebut.



Gambar IV.1 Model konseptual sistem eksperimen terotomatisasi

Dengan struktur yang deterministik, terdokumentasi, dan terotomatisasi, setiap kesimpulan terkait pengaruh persona terhadap penalaran model dapat ditarik secara lebih dapat dipercaya dan dapat dipertanggungjawabkan secara ilmiah.

Tabel IV.1 Analisis komparatif validitas metodologis

Dimensi Analisis	**Sistem Konvensional (Existing)**	**Sistem Usulan (Proposed)**
Pengendalian variabel	Rentan terhadap gangguan input manual; sensitivitas *framing* sulit dikendalikan.	Deterministik; konfigurasi statis dan injeksi otomatis menjamin isolasi variabel independen.
Granularitas data	Hanya menangkap jawaban akhir tanpa metadata.	Telemetri lengkap; menangkap latensi, token, dan jejak penalaran.
Format penyimpanan	Tidak terstruktur; raw text sulit diproses ulang.	Terstruktur (JSON/CSV) dan cocok untuk analisis lanjutan.
Reproduktibilitas	Rendah; parameter eksperimen tidak lengkap terdokumentasi.	Tinggi; seluruh konfigurasi dan kode dapat diaudit dan dijalankan ulang.
Cakupan eksperimen	Terbatas; eksekusi linear menghambat skala eksperimen.	Masif; mendukung ribuan kombinasi melalui *asynchronous execution*.

IV.1.4 Integrasi Persona, Model, dan Benchmark

Bagian ini memaparkan komponen yang membentuk konfigurasi eksperimen, yaitu himpunan persona, himpunan model, dan *benchmark* penalaran yang digunakan.

IV.1.4.1 Benchmark Penalaran

Dua *benchmark* digunakan untuk mengevaluasi dua bentuk penalaran yang berbeda.

GSM8K merupakan kumpulan soal cerita matematika tingkat sekolah menengah yang menguji penalaran numerik bertahap (Cobbe dkk. 2021). Soal-soal GSM8K memiliki struktur jawaban numerik yang jelas sehingga proses evaluasi dapat dilakukan secara deterministik.

MMLU-Redux merupakan versi terkurasi dari MMLU yang memperbaiki ambiguitas format, ketidakkonsistenan pilihan jawaban, dan ketidakseimbangan kualitas antar-subjek (Edinburgh Dataset Analytics Working Group 2024). Benchmark ini digunakan untuk menguji penalaran konseptual lintas domain dengan format pilihan ganda.

Kombinasi GSM8K dan MMLU-Redux memberikan cakupan dua jenis penalaran yang komplementer: numerik prosedural dan konseptual deklaratif.

IV.1.4.2 Himpunan Model

Penelitian ini menggunakan sembilan model bahasa, yang dikelompokkan ke dalam dua kategori berikut:

1. Model komersial: GPT-5, GPT-5 Mini, Claude 4.5 Sonnet, Claude 4.5 Haiku, Gemini 2.5 Flash, Gemini 2.5 Pro;
2. Model publik via OpenRouter: Grok 4.1 Fast, NVIDIA Nemotron-nano-12B-v2-VL, Bert Nebulon Alpha.

Keragaman arsitektur ini memungkinkan analisis komparatif lintas paradigma, mulai dari *frontier models* hingga model publik yang tersedia secara bebas.

IV.1.4.3 Struktur Persona

Persona disusun berdasarkan enam dimensi utama: gender, usia, agama, pekerjaan, kewarganegaraan, dan register bahasa. Kombinasi dimensi tersebut menghasilkan lima belas persona—baik eksplisit maupun implisit—ditambah satu kondisi pengguna netral.

Tabel IV.2 menampilkan daftar lengkap persona yang digunakan.

Tabel IV.2 Daftar user persona untuk kondisi eksperimen

ID	Persona	Mode	Gender	Age Group	Religion	Occupation	Nationality / Register
P1	Implicit male baseline	Implicit	Male	–	–	–	Neutral
P2	Implicit female baseline	Implicit	Female	–	–	–	Neutral
P3	Neutral user	Neutral	–	–	–	–	Neutral
P4	Indonesian Muslim young woman	Explicit	Female	Young adult	Muslim	Healthcare worker	Indonesian / Semi-formal
P5	Indonesian Muslim young man	Implicit	Male	Young adult	Muslim	Healthcare worker	Indonesian / Semi-formal
P6	American middle-aged male	Explicit	Male	Middle-aged	Christian	Engineer	American / Formal
P7	American middle-aged female	Implicit	Female	Middle-aged	Christian	Engineer	American / Formal
P8	Indonesian Gen-Z female	Explicit	Female	Gen-Z	–	Student	Indonesian / Casual-slang
P9	Indonesian Gen-Z male	Implicit	Male	Gen-Z	–	Student	Indonesian / Casual-slang
P10	Middle Eastern young adult male	Explicit	Male	Young adult	Muslim	Engineer	Middle Eastern Arabic / Formal
P11	Middle Eastern young adult female	Implicit	Female	Young adult	Muslim	Student	Middle Eastern Arabic / Formal
P12	American atheist young male	Explicit	Male	Young adult	Atheist	Student	American / Formal
P13	American atheist young female	Implicit	Female	Young adult	Atheist	Student	American / Formal
P14	Indonesian female healthcare worker	Explicit	Female	Young adult	Muslim	Healthcare worker	Indonesian / Semi-formal
P15	Indonesian male healthcare worker	Implicit	Male	Young adult	Muslim	Healthcare worker	Indonesian / Semi-formal

IV.1.4.4 Konfigurasi Eksekusi

Kombinasi lima belas persona dan sembilan model menghasilkan seratus tiga puluh lima konfigurasi eksperimen. Setiap konfigurasi melalui empat tahap eksekusi tetap: pemanasan persona, penetapan konteks percakapan, eksekusi *benchmark*, dan pen-

catatan telemetri. Dengan demikian, setiap variasi keluaran dapat ditelusuri ulang secara langsung kepada kondisi persona yang diberikan.

IV.1.4.5 Contoh Mekanisme Injeksi Persona

Proses injeksi persona dilakukan dengan menyusun *system message* yang mendefinisikan identitas dan karakter pengguna sebelum model menerima stimulus pertanyaan. Dua kategori persona yang digunakan dalam penelitian ini adalah persona eksplisit dan persona implisit.

1. Persona eksplisit.

Pada persona eksplisit, identitas pengguna dirumuskan secara langsung melalui pola deklaratif seperti "your user is ...". Instruksi ini menyebutkan atribut sosial secara jelas—misalnya gender, usia, pekerjaan, atau preferensi gaya bahasa. Contoh injeksi persona eksplisit yang digunakan dalam eksperimen adalah:

> *"Your user is an Indonesian Gen-Z male who works as a junior engineer. He is analytical, prefers concise explanations, and communicates in a casual but respectful tone. Adjust your reasoning structure and tone accordingly when responding."*

Formulasi seperti ini memberikan konteks identitas yang eksplisit sehingga efek persona dapat ditelusuri dengan jelas melalui perubahan pada struktur penalaran dan gaya jawabannya.

2. Persona implisit.

Persona implisit tidak menyatakan identitas sosial secara langsung, tetapi dibangun melalui narasi pengalaman pribadi, emosi, atau gaya tutur tertentu. Model tidak diberi kategori demografis eksplisit; konteks diberikan secara halus, sehingga model perlu menginterpretasikan karakter pengguna melalui sinyal linguistik yang tersirat.

Contoh injeksi persona implisit yang digunakan dalam penelitian ini adalah:

> *"Lately I have been feeling a strange mix of emotional exhaustion and pressure to appear composed, especially when my skin starts acting up unexpectedly. I keep adjusting small routines—like my skincare products—hoping something will finally work. Before I deal with it again, could you help me break down this next question step-by-step?"*

Instruksi seperti ini menghasilkan kerangka persona yang lebih halus dan mendukung analisis terhadap sensitivitas model terhadap gaya bahasa serta cues emosional

yang tidak dinyatakan secara eksplisit.

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

Subbab ini menjelaskan desain arsitektur perangkat lunak yang digunakan untuk merealisasikan *evaluation pipeline* sebagaimana dirumuskan pada Subbab IV.1. Arsitektur pipeline dirancang agar proses evaluasi dapat berjalan secara otomatis, konsisten, dan dapat direproduksi. Pendekatan ini memastikan bahwa setiap kombinasi persona, model, dan *benchmark task* diuji dalam kondisi yang setara dan bebas dari variasi yang tidak diperlukan.

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

IV.2.1 Arsitektur Alur Kerja Sistem

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

1. *Configuration initialization and validation*.
   Tahap ini memuat seluruh konfigurasi sistem, definisi persona, dan *benchmark dataset* ke dalam memori. Validasi struktur data dilakukan untuk memastikan bahwa setiap persona memiliki *system instruction* yang lengkap dan setiap butir tugas memiliki pasangan pertanyaan dan jawaban acuan. Validasi awal ini penting untuk mencegah kesalahan format yang dapat menghentikan proses pada tahap berikutnya.
2. *Prompt construction engine*.
   Pada tahap ini, sistem membentuk dua jenis pesan: *system message* yang berisi identitas persona dan *user message* yang memuat pertanyaan dari benchmark. Penyusunan instruksi dilakukan menggunakan pola yang seragam untuk seluruh iterasi, sehingga setiap model menerima bentuk stimulus yang konsisten. Pendekatan ini menghilangkan variasi yang berasal dari perbedaan penulisan instruksi manual.
3. *Execution manager*.
   Komponen ini mengatur pengiriman permintaan ke model-model bahasa melalui *API interface*. Untuk mengatasi volume permintaan yang besar, *execu-*

tion manager menggunakan pendekatan eksekusi asinkron dengan *I/O concurrency*. Permintaan diatur dalam *task queue* dan dieksekusi dalam kelompok sesuai batas *rate limit*. Strategi ini mempercepat proses pengujian tanpa melampaui kapasitas layanan penyedia model.

4. *Telemetry logger*.
   Komponen terakhir bertanggung jawab menyimpan seluruh respons model dalam format terstruktur, termasuk *model output*, jumlah token, serta *latency*. Data ini digunakan sebagai dasar analisis performa pada bab berikutnya.

Dengan pembagian tersebut, pipeline dapat beroperasi secara modular namun tetap terpadu dalam satu alur pemrosesan.

IV.2.2 Algoritma Orkestrasi dan Konkurensi

Eksperimen dalam penelitian ini melibatkan ribuan kombinasi persona–model–pertanyaan yang menghasilkan volume permintaan API dalam jumlah besar. Eksekusi secara sekuensial tidak praktis karena setiap permintaan memiliki latensi yang bervariasi, sementara penyedia model menerapkan batas *rate limit* yang ketat. Untuk mengatasi hal tersebut, pipeline menggunakan pendekatan eksekusi asinkron berbasis *I/O concurrency*.

Pendekatan ini memungkinkan banyak permintaan dieksekusi secara paralel (hingga batas tertentu), sehingga waktu total dapat ditekan dari kompleksitas $O(N)$ menjadi mendekati $O(N/C)$, dengan C adalah kapasitas konkurensi maksimum. Pipeline membangun sebuah *task queue* yang berisi seluruh pasangan persona–soal, kemudian memprosesnya dalam kelompok (*batch*) sesuai kapasitas konkurensi. Ketika satu batch sedang diproses, sistem dapat menyiapkan batch berikutnya tanpa menunggu seluruh permintaan selesai.

Selain meningkatkan efisiensi waktu, mekanisme ini juga menyediakan ketahanan terhadap kesalahan. Jika terjadi galat seperti *timeout*, *connection reset*, atau `429 Too Many Requests`, pipeline tidak menghentikan seluruh proses. Tugas yang gagal akan dicatat dan dijalankan ulang menggunakan strategi *exponential backoff*, memastikan stabilitas eksekusi jangka panjang.

Algoritma 4.1 berikut mendefinisikan prosedur eksekusi paralel secara formal.

```
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
```

Melalui orkestrasi ini, pipeline mencapai dua tujuan: (1) efisiensi waktu eksekusi yang optimal berkat pemrosesan paralel, dan (2) ketahanan proses melalui penanganan galat adaptif. Dengan demikian, seluruh kombinasi persona–model–benchmark dapat dieksekusi secara konsisten, stabil, dan dapat direproduksi.

IV.2.3 Mekanisme Injeksi Konteks Persona

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

sebelum evaluasi dimulai.

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

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

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

IV.2.4 Mekanisme Toleransi Kesalahan dan Persistensi Status

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

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

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

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

Subbab ini menjelaskan bagaimana rancangan pipeline yang telah disusun pada bagian sebelumnya direalisasikan dalam bentuk organisasi data, struktur direktori, ser-

ta format keluaran yang dihasilkan selama proses eksperimen. Implementasi ini dirancang untuk memastikan bahwa seluruh tahapan pemuatan aset, injeksi konteks persona, pelaksanaan inferensi, dan perekaman hasil berlangsung secara konsisten, terdokumentasi dengan baik, serta mendukung keterulangan eksperimen secara penuh.

IV.3.1 Organisasi Direktori dan Artefak Data

Pipeline dijalankan di atas struktur direktori yang dirancang secara modular untuk memisahkan fungsi pemrosesan dan memudahkan proses audit ilmiah. Empat kelompok artefak utama disusun secara hierarkis sebagai berikut.

1. *Root directory*. Berfungsi sebagai titik masuk eksekusi sistem dan memuat skrip penggerak pipeline beserta utilitas operasional.
2. *Configuration directory*. Menyimpan konfigurasi teknis yang digunakan pipeline, termasuk daftar model, kredensial layanan API, dan parameter eksekusi. Pemisahan direktori ini mendukung aspek keamanan dan memudahkan penggantian parameter tanpa memodifikasi kode utama.
3. *Input assets directory*. Memuat definisi persona serta himpunan benchmark yang telah dinormalisasi. Persona direpresentasikan dalam format terstruktur yang memuat identitas, atribut demografis, dan karakteristik gaya bahasa. Sementara itu, dataset GSM8K dan MMLU Redux dikonversi ke format konsisten untuk memastikan kompatibilitas dengan pipeline.
4. *Results directory*. Menyimpan keseluruhan keluaran eksperimen yang mencakup log granular pada tingkat per butir soal, tabel hasil, serta agregasi lintas persona dan lintas model. Struktur ini memudahkan proses penelusuran kembali bagi keperluan analisis.

Pemilahan direktori ini memastikan bahwa seluruh artefak eksperimen terdokumentasi secara terstruktur dan mudah direplikasi.

IV.3.2 Subsistem Perangkat Lunak dan Alur Transformasi Data

Pipeline terdiri atas empat subsistem utama yang bekerja secara berurutan dalam mengelola eksekusi eksperimen:

1. *Execution orchestration subsystem*. Subsistem ini membentuk *task queue* yang memuat seluruh kombinasi model, persona, dan pertanyaan. Orkestrasi ini memastikan determinisme dan menghindari variasi eksekusi akibat intervensi manual.
2. *Model communication subsystem*. Bertugas melakukan konstruksi instruk-

si, mengirimkan permintaan ke layanan model, menangani kode galat, serta menegakkan batas layanan seperti *rate limit*. Seluruh komunikasi dilakukan menggunakan protokol API yang distandardisasi.

3. *Monitoring subsystem*. Menyediakan mekanisme *checkpointing* sehingga eksekusi dapat dilanjutkan tanpa kehilangan progres apabila terjadi gangguan jaringan atau penghentian proses secara tidak terduga. Hal ini memastikan konsistensi eksekusi dan mengurangi risiko duplikasi.
4. *Analysis subsystem*. Mengolah log mentah menjadi tabel terstruktur dan menghitung metrik utama seperti akurasi, penggunaan token, dan latensi. Modul ini menghasilkan keluaran agregasi yang digunakan dalam tahap analisis pada Bab V.

Alur transformasi data berlangsung dari log granular menuju tabel pemetaan kemudian agregasi lintas model, sehingga memungkinkan analisis kuantitatif yang komprehensif.

IV.3.3 Representasi Persona dan Mekanisme Injeksi Konteks

Persona direpresentasikan dalam format terstruktur yang memuat identitas demografis, atribut gaya bahasa, serta narasi yang relevan. Representasi tersebut kemudian dikonversi menjadi *system instruction* yang ditempatkan pada segmen instruksi sistem saat permintaan dikirimkan ke model.

Injeksi konteks persona dilakukan satu kali melalui dua tahap:

1. *Persona grounding*. Tahap ini menanamkan identitas dan karakteristik gaya bahasa persona secara eksplisit atau implisit pada konteks model.
2. *Warm up interaction*. Dilakukan satu interaksi pemanasan untuk menstabilkan perilaku model sehingga respons pada tahap berikutnya mengikuti karakter persona secara konsisten.

Setelah kedua tahap tersebut selesai, pipeline mengirim seluruh pertanyaan GSM8K dan MMLU Redux tanpa mengulang injeksi persona. Dengan demikian, kondisi kognitif model dijaga agar tetap setara di seluruh siklus inferensi.

IV.3.4 Contoh Struktur Log Inferensi

Untuk menjaga transparansi dan keterulangan eksperimen, pipeline mencatat setiap interaksi dengan model dalam bentuk log terstruktur. Log ini memuat informasi mengenai konfigurasi eksekusi, isi jawaban model, serta telemetri penggunaan token. Cuplikan pada Kode IV.3.4 menunjukkan contoh keluaran untuk model yang tidak menyediakan *reasoning trace*.

Kode IV.1 Contoh log inferensi tanpa reasoning trace

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

Pada model tertentu, layanan juga menyediakan informasi tambahan mengenai proses penalaran internal yang digunakan untuk menghasilkan jawaban akhir. Informasi ini direkam sebagai *reasoning trace* dan disimpan terpisah dari konten jawaban. Kode IV.3.4 memperlihatkan contoh log untuk model yang menyediakan *reasoning trace* beserta jumlah token yang digunakan pada bagian tersebut.

Kode IV.2 Contoh log inferensi dengan reasoning trace

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

Kedua contoh tersebut menggambarkan bagaimana pipeline menangkap tidak hanya jawaban akhir, tetapi juga struktur penalaran dan sumber daya komputasi yang digunakan oleh model. Informasi ini menjadi dasar analisis lebih lanjut mengenai perbedaan perilaku antar model dan antar persona.

IV.3.5 Ringkasan Hasil Eksperimen

Ringkasan performa lintas model dan persona ditampilkan pada Tabel IV.3. Tabel ini memberikan gambaran umum mengenai tingkat akurasi dan beban komputasi untuk setiap konfigurasi, dan digunakan sebagai dasar analisis pada Bab V.

Tabel IV.3 Ringkasan Hasil Eksperimen GSM8K untuk Seluruh Model dan Persona

Model	Persona	Total Q	Correct	Accuracy (%)	Total Tokens
Bert Nebulon Alpha	man_implicits	610	593	97.21	285250
Bert Nebulon Alpha	woman_implicits	641	627	97.26	335208
Grok 4.1 Fast	man_implicits	1315	1242	94.45	1325229
Grok 4.1 Fast	woman_implicits	1316	1254	95.36	1422736
Nvidia Nemotron 12B v2 VL	man_implicits	1305	1224	93.79	1156049
Nvidia Nemotron 12B v2 VL	woman_implicits	1315	1248	94.98	1986284

Tabel tersebut menjadi dasar perbandingan antar model pada bab evaluasi, termasuk pengaruh persona terhadap akurasi dan kompleksitas respons.

BAB V

RENCANA SELANJUTNYA

V.1 Rencana Implementasi dan Estimasi Biaya

Rencana implementasi pada tahap berikutnya adalah menjalankan kembali *evaluation pipeline* yang telah dijelaskan pada Bab IV dengan cakupan penuh, mencakup sembilan model bahasa, dua *benchmark* penalaran (GSM8K dan MMLU-Redux), serta lima belas *user persona* (implisit, eksplisit, dan netral). Bagian ini merumuskan langkah implementasi teknis, asumsi kebutuhan token, serta estimasi biaya penggunaan API berdasarkan harga resmi masing-masing model pada platform OpenRouter[1]

Estimasi dilakukan menggunakan kurs konstan 1 USD = Rp16.000.

V.1.1 Rencana Implementasi Eksperimen

Implementasi eksperimen direncanakan mengikuti enam langkah utama berikut.

1. Persiapan aset data.
   Sistem memuat berkas definisi 15 persona, korpus GSM8K (split *test*), MMLU-Redux (20 subjek), kredensial API, dan konfigurasi model. Struktur direktori dan modul pemrosesan mengikuti rancangan pada Subbab **??**.
2. Inisialisasi dan *warm-up* persona.
   Setiap model menerima satu pesan awal untuk menanamkan konteks persona sebelum mengerjakan soal pertama. Tahap ini juga berfungsi sebagai *sanity check* bahwa model mengikuti identitas dan gaya bahasa persona.
3. Eksekusi eksperimen utama.
   Setiap kombinasi model–persona menjalankan seluruh soal GSM8K dan MMLU-

1. *OpenAI GPT-5 Mini Pricing*; *Anthropic Claude Haiku 4.5 Pricing*; *Google Gemini 2.5 Flash Pricing*; *DeepSeek V3.2 Pricing*; *NVIDIA Llama 3.3 Nemotron Super 49B V1.5 Pricing*; *Google Gemma 3n 4B Pricing*.

Redux menggunakan mekanisme injeksi pesan berbasis peran: persona pada *system message* dan soal pada *user message*. Setiap respons diharuskan mencakup penalaran langkah demi langkah.

4. Pencatatan log granular.
   Setiap respons disimpan sebagai log JSON yang memuat isi *prompt*, jawaban mentah, *token usage*, dan *latency*.
5. Agregasi dan validasi hasil.
   Data log diubah menjadi CSV agregat yang berisi akurasi, rata-rata latensi, dan konsumsi token. Validasi tambahan mencakup pemeriksaan pola jawaban dan konsistensi jumlah entri.
6. Penanganan kegagalan.
   Kegagalan akibat *timeout* atau batas *rate limit* ditangani menggunakan *retry* dengan *exponential backoff*, sesuai mekanisme pada Bab IV. Dengan demikian, kegagalan sebagian tidak mengganggu keseluruhan eksperimen.

V.1.2 Himpunan Model dan Skenario Eksekusi

Eksperimen ini menggunakan sembilan model dengan rincian sebagai berikut.

1. Enam model berbayar (via OpenRouter):
   (a) openai/gpt-5-mini
   (b) anthropic/claude-haiku-4.5
   (c) google/gemini-2.5-flash
   (d) deepseek/deepseek-v3.2
   (e) nvidia/llama-3.3-nemotron-super-49b-v1.5
   (f) google/gemma-3n-e4b-it
2. Tiga model yang pada saat perancangan tersedia sebagai *free-tier*:
   (a) xai/grok-4.1-fast
   (b) nvidia/nemotron-nano-12b-v2-vl
   (c) openrouter/bert-nebulon-alpha

Seluruh sembilan model dijalankan pada konfigurasi penuh: dua *benchmark* dan lima belas persona. Namun, estimasi biaya hanya dihitung untuk enam model berbayar.

V.1.3 Asumsi Jumlah Soal dan Kebutuhan Token

Kebutuhan token dihitung berdasarkan dua sumber utama: GSM8K (1319 soal) dan MMLU-Redux (2000 soal). Pada kedua *benchmark*, model diarahkan untuk memberikan penalaran lengkap sebelum jawaban akhir, sehingga konsumsi token per

soal diharapkan berada pada kisaran yang relatif tinggi.

1. GSM8K.
   Total token per persona per model diestimasikan sebagai:

$$T_{\text{GSM8K}} \approx 1319 \times 1200 = 1{,}582{,}800 \text{ token}.$$

2. MMLU-Redux.
   Total token per persona per model diestimasikan sebagai:

$$T_{\text{MMLU}} \approx 2000 \times 1200 = 2{,}400{,}000 \text{ token}.$$

Total token inti per persona diperoleh dari penjumlahan keduanya:

$$T_{\text{base, persona}} = 1{,}582{,}800 + 2{,}400{,}000 = 3{,}982{,}800.$$

Untuk mengakomodasi *warm-up* dan *retry*, digunakan faktor overhead 20%:

$$T_{\text{persona}} \approx 1.2 \times 3{,}982{,}800 = 4{,}779{,}360.$$

Sehingga total token per model untuk 15 persona adalah:

$$T_{\text{model}} \approx 15 \times 4{,}779{,}360 = 71{,}690{,}400 \approx 71{,}7 \times 10^6.$$

Komposisi token diasumsikan:

$$T_{\text{in}} = 0.4T_{\text{model}}, \qquad T_{\text{out}} = 0.6T_{\text{model}}.$$

V.1.4 Estimasi Biaya per Model

Harga token per model mengacu pada dokumentasi OpenRouter(*OpenAI GPT-5 Mini Pricing*; *Anthropic Claude Haiku 4.5 Pricing*; *Google Gemini 2.5 Flash Pricing*; *DeepSeek V3.2 Pricing*; *NVIDIA Llama 3.3 Nemotron Super 49B V1.5 Pricing*; *Google Gemma 3n 4B Pricing*). Biaya untuk model ke-m dihitung dengan rumus:

$$\text{cost}_m = p_{\text{in},m} \times \frac{T_{\text{in}}}{10^6} + p_{\text{out},m} \times \frac{T_{\text{out}}}{10^6},$$

dengan $p_{\text{in},m}$ dan $p_{\text{out},m}$ adalah harga per satu juta token untuk *input* dan *output*.

Estimasi berikut menggunakan kurs Rp 16.000 per USD dan total token $T_{\text{model}} \approx 71{,}7 \times 10^6$.

Tabel V.1 Estimasi biaya enam model berbayar untuk konfigurasi penuh 15 persona

Model	Total token T_{model}	Biaya (USD)	Biaya (Rp)
openai/gpt-5-mini	$\approx 71{,}7 \times 10^6$	93.20	$\approx 1{,}491{,}000$
anthropic/claude-haiku-4.5	$\approx 71{,}7 \times 10^6$	243.75	$\approx 3{,}900{,}000$
google/gemini-2.5-flash	$\approx 71{,}7 \times 10^6$	116.14	$\approx 1{,}858{,}000$
deepseek/deepseek-v3.2	$\approx 71{,}7 \times 10^6$	24.95	$\approx 399{,}000$
nvidia/llama-3.3-nemotron-super-49b-v1.5	$\approx 71{,}7 \times 10^6$	20.07	$\approx 321{,}000$
google/gemma-3n-e4b-it	$\approx 71{,}7 \times 10^6$	2.29	$\approx 37{,}000$
Total enam model berbayar	–	500.40	$\approx 8{,}006{,}000$

Tiga model lain yang tersedia sebagai *free-tier* (grok-4.1-fast, nemotron-nano-12b-v2-vl, dan bert-nebulon-alpha) diperkirakan mengonsumsi token serupa tetapi tidak menimbulkan biaya finansial langsung. Status *free-tier* tersebut tetap harus diverifikasi kembali sebelum eksperimen akhir dijalankan.

Dengan demikian, estimasi total biaya finansial untuk menjalankan seluruh eksperimen multi-model, multi-persona, dan dua *benchmark* penalaran adalah sekitar 500,40 USD atau kurang lebih 8 juta rupiah. Angka ini bersifat konservatif karena telah memasukkan biaya *warm-up* dan *retry*, sehingga realisasi biaya dapat lebih rendah apabila konsumsi token aktual per soal ternyata lebih kecil dari asumsi yang digunakan dalam perhitungan ini.

V.2 Desain Pengujian dan Evaluasi

Desain pengujian pada tahap berikut disusun untuk memastikan bahwa seluruh hasil eksperimen dapat diverifikasi, divalidasi, dan direplikasi. Struktur pengujian memanfaatkan artefak log granular, telemetry penggunaan token, serta pemeriksaan konsistensi yang telah ditanamkan dalam pipeline pada Bab IV.

1. Verifikasi konsistensi eksekusi.
   Verifikasi dilakukan untuk memastikan bahwa setiap model menerima stimulus yang identik pada setiap soal dan persona, sehingga variasi respons dapat dikaitkan langsung dengan perbedaan persona atau arsitektur model.
   (i) Konsistensi konstruksi prompt.
   Pemeriksaan dilakukan untuk memastikan bahwa struktur persona pada *sys-*

tem message dan isi soal pada *user message* identik pada seluruh eksekusi. Setiap variasi kecil seperti pergeseran tanda baca atau perubahan format dapat mengubah jalur penalaran model, sehingga pemeriksaan dilakukan secara programatik pada log JSON.

(ii) Kesesuaian urutan eksekusi.

Pemeriksaan dilakukan dengan mencocokkan indeks interaksi, nomor soal, dan urutan persona pada seluruh berkas log untuk memastikan bahwa sistem menjalankan eksperimen sesuai konfigurasi yang direncanakan.

(iii) Keberhasilan tahap warm-up.

Tahap warm-up diverifikasi dengan menilai apakah respons awal model mengikuti identitas dan gaya bahasa persona. Kegagalan tahap ini dicatat sebagai anomali dan disertai eksekusi ulang sebelum proses utama dimulai.

2. Validasi keluaran model.

   Validasi keluaran bertujuan memastikan bahwa jawaban model berada dalam format yang sesuai untuk dievaluasi. Pendekatan validasi dibedakan untuk GSM8K dan MMLU-Redux.

   (i) Validasi GSM8K.

   Model harus memberikan jawaban numerik akhir yang dapat diekstraksi secara deterministik. Selain itu, respons harus mencakup penalaran langkah demi langkah sebelum menyatakan jawaban akhir.

   (ii) Validasi MMLU-Redux.

   Model harus memberikan pilihan jawaban dalam format A, B, C, atau D. Meskipun merupakan soal pilihan ganda, model tetap diminta menjelaskan penalaran sebelum memilih opsi, sehingga respons memiliki struktur yang konsisten.

   (iii) Pemeriksaan konsistensi format respons.

   Pemeriksaan mencakup panjang respons, struktur teks, keberadaan penalaran, serta keterbacaan sehingga setiap respons dapat diproses ulang tanpa kesalahan parsing.

3. Evaluasi kuantitatif.

   Evaluasi kuantitatif dilakukan untuk mengukur dampak persona terhadap performa model pada dua benchmark.

   (i) Akurasi jawaban.

   Akurasi dihitung dengan membandingkan jawaban akhir yang diekstraksi terhadap ground truth. Penghitungan dilakukan pada tabel agregasi hasil.

   (ii) Konsumsi token.

   Evaluasi melibatkan token input, token output, dan token penalaran sebagai

indikator beban komputasi dan kecenderungan verbosity model di bawah persona tertentu.

(iii) Latensi eksekusi.

Latensi diambil dari metadata waktu pada log JSON untuk menilai stabilitas waktu respons model ketika menangani beban besar dan variasi persona.

V.3 Analisis Risiko dan Mitigasi

Pelaksanaan eksperimen pada lingkungan multi-model dan multi-persona menimbulkan sejumlah risiko metodologis dan operasional yang perlu dikelola secara sistematis agar integritas penelitian tetap terjaga. Risiko-risiko tersebut mencakup aspek reliabilitas penggunaan API, kestabilan keluaran model, konsistensi proses penalaran, menjaga ketepatan pesan sepanjang percakapan, serta akurasi proses agregasi data. Selain itu, penelitian sebelumnya menunjukkan bahwa konsistensi LLM dapat menurun pada evaluasi berskala besar dan bahwa penalaran model dapat terpengaruh oleh faktor-faktor non-linguistik yang tidak terkontrol. Berdasarkan temuan tersebut, bagian ini menguraikan tiga kategori risiko utama serta strategi mitigasinya.

1. Risiko kegagalan pemanggilan API.

   Risiko ini mencakup galat seperti *timeout*, gangguan koneksi, dan pembatasan layanan (rate limit). Kegagalan ini berpotensi menyebabkan hilangnya sebagian data atau ketidaksinkronan indeks percobaan.

   (i) Mitigasi dilakukan melalui mekanisme *retry* adaptif berbasis *exponential backoff*, sesuai praktik standar pada sistem terdistribusi.

   (ii) Seluruh kegagalan direkam dalam log terpisah untuk memastikan keterlacakan sehingga perbaikan atau pengulangan dapat dilakukan secara selektif.

   (iii) Tingkat konkurensi dijalankan secara otomatis ketika sistem mendeteksi peningkatan laju galat, guna menjaga stabilitas kapasitas layanan.

2. Risiko lonjakan konsumsi token.

   LLM sering menghasilkan keluaran yang lebih panjang daripada yang diinstruksikan, terutama ketika diminta memberikan penalaran langkah demi langkah. Fenomena ini berdampak langsung pada biaya dan durasi eksperimen.

   (i) Sistem membatasi panjang keluaran dengan parameter *maximum completion length* untuk mencegah respons berlebihan.

   (ii) Validasi awal dijalankan secara berkala untuk memantau rata-rata konsumsi token per soal.

   (iii) Persona yang terbukti memicu keluaran terlalu panjang dilakukan pe-

nyesuaian instruksi secara minimal untuk mengendalikan panjang teks tanpa mengubah maksud identitas sosial.

3. Risiko penyimpanan dan konsistensi log.
   Volume log yang besar berpotensi menimbulkan risiko korupsi berkas dan ketidakcocokan antara indeks model, persona, dan soal.
   (i) Setiap respons disimpan dalam format terstruktur (JSON) dengan skema tetap.
   (ii) Proses agregasi mengadopsi pemeriksaan konsistensi silang antara jumlah entri dan indeks soal.
   (iii) Mekanisme *checkpointing* diterapkan untuk menghindari kehilangan data apabila eksekusi terhenti di tengah proses.

DAFTAR PUSTAKA

Anthropic Claude Haiku 4.5 Pricing. https://openrouter.ai/anthropic/claude-haiku-4.5. Diakses 2025.

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.

DeepSeek V3.2 Pricing. https://openrouter.ai/deepseek/deepseek-v3.2. Diakses 2025.

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

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

Google Gemini 2.5 Flash Pricing. https://openrouter.ai/google/gemini-2.5-flash. Diakses 2025.

Google Gemma 3n 4B Pricing. https://openrouter.ai/google/gemma-3n-e4b-it. Diakses 2025.

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

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

Liang, P., R. Bommasani, dkk. 2023. "Holistic Evaluation of Language Models". *arXiv preprint arXiv:2211.09110,* https://arxiv.org/abs/2211.09110.

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

NVIDIA Llama 3.3 Nemotron Super 49B V1.5 Pricing. https://openrouter.ai/nvidia/llama-3.3-nemotron-super-49b-v1.5. Diakses 2025.

OpenAI GPT-5 Mini Pricing. https://openrouter.ai/openai/gpt-5-mini. Diakses 2025.

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.