Jurnal Informatika: Jurnal pengembangan IT

Vol. 10, No. 4, 2025
ISSN : 2477-5126, DOI:10.30591/jpit.v10i4.9047

Evaluasi Gemini Flash pada Ekstraksi Jadwal Skripsi Terstruktur dan Tidak Terstruktur

Galih Hermawan [1], Ednawati Rainarli [2]

[1]Program Studi Teknik Informatika, Universitas Komputer Indonesia, Jl. Dipati Ukur No. 112-114, Bandung, 40132, Indonesia
[2]Program Studi Teknik Robotika dan Kecerdasan Buatan, Universitas Komputer Indonesia, Jl. Dipati Ukur No. 112-114, Bandung, 40132, Indonesia

Info Artikel

Riwayat Artikel:

Received 2025-06-19
Revised 2025-10-20
Accepted 2025-10-21

Corresponding Author:

Galih Hermawan

Email:

galih.hermawan@email.unikom.ac.id

This is an open access article under the CC BY 4.0 license.

Abstract *– The administration of thesis seminar and defense scheduling is often hampered by unstructured PDF formats, which increases manual workload and the risk of errors. This study aims to evaluate and compare the performance of three Gemini Flash model variants, namely Gemini 2.0 Flash-Lite, Gemini 2.0 Flash, and Gemini 2.5 Flash Preview, in automating schedule information extraction using a zero-shot prompting approach. The dataset consists of 87 PDF files containing thesis seminar and defense schedules (588 entries) from the 2023/2024 academic year, alongside 200 question scenarios executed in two different context formats: raw extracted text (TXT) and structured JSON data. Performance evaluation metrics include Precision, Recall, F1-score, Exact-Match, and inference latency per request. Experimental results indicate that Gemini 2.5 Flash Preview achieves average F1-scores above 0.98 in both contexts with approximately 3.9 seconds latency. Conversely, smaller-capacity variants (Gemini 2.0 Flash and Flash-Lite) showed more significant performance gains using the JSON format compared to raw text, especially on complex question types such as multi-attribute filtering and list retrieval. Through error analysis, the primary challenge identified was tasks requiring numeric aggregation and determination of superlative values, accounting for approximately 78% of total extraction failures, particularly for lightweight models. A paired t-test indicated no statistically significant difference between the two context formats (average F1 difference = 0.0077; p=0.48). This study recommends the use of explicit numeric prompting or rule-based post-processing when employing lightweight models to significantly improve the accuracy of academic schedule information extraction.*

Keywords: *Gemini Flash, Information Extraction, Large Language Models, Thesis Scheduling, Zero-Shot Prompting.*

Abstrak *– Administrasi penjadwalan skripsi sering terhambat oleh format PDF tidak terstruktur, sehingga meningkatkan beban kerja manual dan risiko kesalahan. Penelitian ini bertujuan mengevaluasi dan membandingkan kinerja tiga varian model Gemini Flash, yaitu Gemini 2.0 Flash-Lite, Gemini 2.0 Flash, dan Gemini 2.5 Flash Preview, dalam otomatisasi ekstraksi informasi jadwal dengan pendekatan zero-shot prompting. Dataset mencakup 87 berkas PDF jadwal seminar dan sidang skripsi (588 entri) tahun akademik 2023/2024, serta 200 skenario pertanyaan yang dijalankan pada dua jenis konteks data: teks mentah hasil ekstraksi langsung (TXT) dan data terstruktur berbasis JSON. Evaluasi kinerja dilakukan menggunakan metrik Precision, Recall, F1-score, Exact-Match, serta latensi inferensi per permintaan. Hasil eksperimen menunjukkan Gemini 2.5 Flash Preview mencapai skor F1 rata-rata di atas 0,98 pada kedua format konteks dengan latensi sekitar 3,9 detik. Sebaliknya, varian dengan kapasitas lebih kecil (Gemini 2.0 Flash dan Flash-Lite) memperoleh peningkatan kinerja yang lebih signifikan dengan format JSON dibandingkan teks mentah, terutama untuk pertanyaan kompleks seperti filter multi-atribut dan list retrieval. Tantangan utama yang teridentifikasi melalui analisis kesalahan adalah pada tugas yang membutuhkan agregasi numerik dan penentuan nilai ekstrem, yang menyumbang sekitar 78% dari total kegagalan ekstraksi, terutama pada model berkapasitas ringan. Uji paired t-test memperlihatkan tidak adanya perbedaan signifikan antara kinerja kedua format konteks (selisih F1 rata-rata 0,0077; p=0,48). Penelitian ini merekomendasikan penggunaan prompt numerik eksplisit atau pascaproses berbasis aturan dalam penerapan model ringan untuk meningkatkan akurasi ekstraksi informasi jadwal akademik secara signifikan.*

Kata Kunci*: Ekstraksi Informasi, Gemini Flash, Jadwal Skripsi, Model Bahasa Besar, Zero-Shot Prompting.*

I. PENDAHULUAN

Kegiatan seminar dan sidang skripsi merupakan tahapan penentu kelulusan yang lazim ditemui di berbagai program studi informatika di perguruan tinggi. Meskipun terminologi, urutan prosedur, dan bentuk penilaian dapat bervariasi antar kampus, esensinya tetap sama: mahasiswa mempresentasikan kemajuan riset (seminar) dan mempertahankannya di hadapan penguji (sidang). Di Program Studi Teknik Informatika, Universitas Komputer Indonesia (UNIKOM), panitia akademik membuka pendaftaran, memverifikasi berkas, dan menyusun jadwal pelaksanaan. Menjelang setiap pekan pelaksanaan, jadwal diunggah ke repositori Google Drive dalam bentuk berkas PDF bertabel yang memuat nomor induk mahasiswa (NIM), nama mahasiswa, judul, waktu, ruang, serta nama dosen

pembimbing dan penguji. Praktik distribusi PDF ini merupakan prosedur administratif internal; institusi lain dapat memakai kanal atau format berbeda, tetapi pola penyebaran jadwal statis serupa tetap berpotensi menimbulkan beban informasi.

Penelitian ini memanfaatkan arsip jadwal semester Ganjil dan Genap 2023/2024 yang dikelola Program Studi Teknik Informatika UNIKOM. Dataset tersebut berisi 87 berkas PDF dengan total 588 entri jadwal, terdiri atas 157 entri pada semester Ganjil dan 431 entri pada semester Genap. Distribusi dokumen setiap pekan menimbulkan *information overload* [1] sekaligus menambah beban kerja administratif dosen [2], karena mereka harus membuka banyak fail secara manual untuk menemukan jadwal relevan, sehingga meningkatkan risiko kesalahan administratif seperti salah ruangan atau waktu pelaksanaan. Laporan *Organisation for Economic Co operation and Development* (OECD), sebuah organisasi ekonomi antar pemerintah, menunjukkan bahwa beban administratif semacam ini menurunkan efisiensi kegiatan akademik [3], sedangkan studi Diallo dan Tudose [4] menekankan pentingnya optimasi penjadwalan untuk meningkatkan efektivitas institusi pendidikan tinggi.

Perkembangan model bahasa besar atau *Large Language Models* (LLM), khususnya setelah keberhasilan pendekatan *few-shot learning* oleh GPT-3 [5], telah mendorong banyak penelitian terkait ekstraksi informasi dari dokumen kompleks. Xu dkk. [6], Grothey dkk. [7], dan Ntinopoulos dkk. [8] secara khusus menunjukkan kemampuan LLM dalam ekstraksi informasi dari dokumen medis dan klinis [9]. Jobson dan Li [10] secara spesifik menyelidiki potensi LLM dalam manajemen penjadwalan. Tantangan umum yang ditemukan dalam implementasi LLM mencakup tokenisasi dan sensitivitas *subword*, seperti dilaporkan Chai dkk. [11] dan Chang & Bergen [12]. Studi lainnya memperlihatkan keberhasilan LLM dalam ekstraksi tabel dokumen finansial [13]. Meski demikian, evaluasi pemanfaatan LLM khusus untuk ekstraksi jadwal akademik berbahasa Indonesia masih terbatas.

Penelitian ini mengeksplorasi potensi tiga varian Gemini Flash (Gemini 2.0 Flash-Lite, Gemini 2.0 Flash, Gemini 2.5 Flash Preview [14], [15], [16]) dalam otomatisasi ekstraksi informasi dari dokumen jadwal PDF. Kontribusi utama penelitian ini meliputi: (1) penyediaan korpus publik 588 entri jadwal skripsi beserta ground-truth dalam format JSONL, (2) protokol evaluasi *record-level* dengan metrik *Precision*, *Recall*, *F1-score*, dan *Exact-Match*, serta (3) analisis agregat pola kesalahan ekstraksi. Guna memperoleh cakupan evaluasi yang komprehensif, 200 pertanyaan uji dikelompokkan ke dalam enam tipe tugas: *lookup* nama (K-1), *lookup* NIM (K-2), filter multi-atribut (K-3), *list retrieval* (K-4), agregasi numerik (K-5), dan superlatif deterministik (K-6). Pengelompokan ini mengadaptasi kategori tugas (*task-archetype*) yang lazim pada *benchmark question answering* (QA) dan *information retrieval* (IR) seperti SQuAD [17], *Fact, Fetch* & *Reason* [18], *Lost in the Middle* [19], DROP [20], dan SYGMA [21].

Secara khusus, penelitian ini bertujuan untuk mengukur akurasi ekstraksi entitas dan atribut jadwal skripsi dari masing-masing varian Gemini Flash yang diuji; mengevaluasi dampak format konteks, yaitu teks mentah (TXT) dibandingkan JSON terstruktur, terhadap kinerja ekstraksi; mengidentifikasi pola kegagalan atau kesalahan dominan dalam tugas ekstraksi jadwal; serta mengevaluasi pertukaran (*trade-off*) antara akurasi dan latensi inferensi ketiga varian model ketika dipertimbangkan dalam integrasi sistem manajemen akademik. Bagian berikut dari makalah ini disusun sebagai berikut: Bagian II menguraikan metodologi penelitian, termasuk dataset dan skema eksperimen. Bagian III memaparkan hasil evaluasi dan pembahasan performa LLM. Bagian IV menyimpulkan hasil penelitian serta memberikan rekomendasi untuk penelitian lanjutan.

II. METODE

Pada bab ini akan dijelaskan langkah-langkah penelitian yang dilakukan. Proses pengumpulan data masukan, model LLM, metrik, dan alur pengujian yang digunakan untuk mengukur performansi LLM yang telah diadaptasi.

A. Data Masukan

Ada tiga data masukan yang dibutuhkan untuk evaluasi LLM, yaitu data konteks (data terstruktur dan tidak terstruktur), data daftar pertanyaan yang akan diuji beserta jawabannya, dan *prompt* yang digunakan untuk membuat model dapat beradaptasi.

JADWAL SEMINAR
TAHUN AKADEMIK 2023/2024 SEMESTER GENAP
JUMAT, 5-JULI-2024

NO	NIM	NAMA	JUDUL	REVIEWER & PEMBIMBINGS	WAKTU	RUANG
1	10120758	FAJAR LAKSANA	Sistem Penentuan Supplier Terbaik Menggunakan Metode Analytical Hierarchy Process (AHP) di Casanirah	Sufaatin, S.T., M.Kom. Gentisya Tri Mardiani, S.Kom., M.Kom	08.00-09.15	R6010
2	10120233	ARDIAN JULIANTO SAGALA	Situs Web Pemantauan Pesawat Terbang Rute Bandara Indonesia	Chrismikha Haryanto, S.Kom.,M.Kom. Andri Heryandi, S.T., M.T.	08.00-09.15	R6028
3	10120784	DENISA ALYAA PUTRI	Sistem Penentuan Jumlah dan Waktu Pembelian Barang Menggunakan Metode Economic Order Quantity (EOQ) di PT. Hartelindo Telco Utama	Gentisya Tri Mardiani, S.Kom., M.Kom Sri Nurhayati, S.Si., M.T.	09.15-10.30	R6010
4	10120068	ALBERIANSYAH	Implementasi Sistem Absensi Berbasis GPS, WIFI Positioning dan Sidik Jari untuk Peningkatan Akurasi dan Keamanan Absensi	Andri Heryandi, S.T., M.T. Bobi Kurniawan Soegoto, ST.,M.Kom	09.15-10.30	R6028
5	10120103	MUHAMMAD RUKKI AFSA JATNIKA	Pembangunan Telegram Chatbot "U-Learning" Berbasis Artificial Inteligence sebagai Asistem Pembelajaran Mandiri Siswa	Chrismikha Haryanto, S.Kom.,M.Kom. Irfan Maliki, S.T., M.T.	10.30-11.45	R6028

Gambar 1. Contoh isi fail jadwal seminar dalam format PDF

Gambar 1 adalah contoh fail jadwal dengan format PDF. Adapun informasi yang terdapat dalam fail jadwal seminar adalah berupa hari, tanggal, jam, NIM, dan nama mahasiswa, nama dosen pembimbing/penguji serta ruang pelaksanaan seminar/sidang.

Fail PDF mentah diekstraksi menjadi teks mentah per halaman menggunakan *library* PyMuPDF (fitz) melalui fungsi *get_text*(). Proses ini mempertahankan tata letak tabel dokumen asli, tetapi tidak menghasilkan struktur data eksplisit. Karena setiap berkas PDF merupakan digital asli (*born-digital*), ekstraksi teks dilakukan tanpa memerlukan teknologi OCR. Hasil ekstraksi tersebut kemudian digunakan dalam dua bentuk konteks, yaitu teks mentah sebagai data tidak terstruktur dan JSON sebagai data terstruktur, sebagaimana ditunjukkan pada Gambar 2.

```
==================================================
Senin, 27 November 2023.pdf
==================================================
JADWAL SEMINAR
NO
NIM
NAMA
JUDUL
REVIEWER & PEMBIMBINGS
WAKTU
RUANG
TAHUN AKADEMIK 2023/2024 SEMESTER GANJIL
SENIN, 27-NOVEMBER-2023
WILDAN LUTFI ANNANTA
1
10118143
Rancang Bangun Kamera Dinamis sebagai Monitoring
Pertumbuhan Tanaman berbasis Internet of Things (Iot)
Dedeng Hirawan, S.Kom., M.Kom
13.00-14.15
R6010
Didit Andri Jatmiko, S.Kom., M.T.
```

```
{
  "semester": "Ganjil",
  "tahun_akademik": "2023/2024",
  "pekan": 1,
  "jenis": "Seminar",
  "jadwal_harian": [
    {
      "tanggal": "2023-11-27",
      "hari": "Senin",
      "jadwal": [
        {
          "nim": "10118143",
          "nama": "Wildan Lutfi Annanta",
          "judul": "Rancang Bangun Kamera Dinamis sebagai
          Monitoring Pertumbuhan Tanaman berbasis Internet of
          Things (IoT)",
          "waktu_mulai": "13:00",
          "waktu_selesai": "14:15",
          "ruang": "R6010",
          "dosen": [
            {
              "nama": "Dedeng Hirawan, S.Kom., M.Kom.",
              "peran": "Pembimbing"
            },
            {
              "nama": "Didit Andri Jatmiko, S.Kom., M.T.",
              "peran": "Reviewer"
            }
          ]
        },
```

(a) (b)

Gambar 2. Format Data Konteks pada Eksperimen: (a) Teks Mentah, (b) JSON Terstruktur

Gambar 2(a) menampilkan contoh luaran TXT hasil ekstraksi, sedangkan Gambar 2(b) memperlihatkan versi JSON terstruktur yang dibangun secara semi-otomatis dengan atribut seperti tanggal, waktu, ruang, NIM, nama mahasiswa, judul skripsi, serta daftar dosen pembimbing dan penguji. Fail TXT digunakan sebagai masukan LLM dalam bentuk data tak terstruktur, sedangkan fail JSON terstruktur berfungsi sebagai masukan terstruktur. Seluruh berkas PDF mentah, hasil konversi ke TXT dan JSON terstruktur, bersama dataset *skenario.jsonl*, *ground-truth.jsonl*, dan berkas hasil evaluasi dipublikasikan di repositori GitHub untuk mendukung reproduksibilitas [22].

Untuk menguji kinerja dari model LLM dibuat skenario pengujian menggunakan 200 pertanyaan terkait jadwal sidang/seminar skripsi. Pertanyaan tersebut dikelompokkan berdasarkan dimensi yang ingin diukur dari model LLM. Skenario pengujian dalam penelitian ini didefinisikan sebagai unit observasi utama yang merepresentasikan satu instruksi atau pertanyaan berbasis bahasa alami, dengan tujuan mengevaluasi kemampuan model LLM dalam mengekstraksi informasi spesifik dari data jadwal seminar dan sidang skripsi. Perancangan skenario didasarkan pada kebutuhan nyata dosen, mahasiswa, maupun admin program studi terhadap akses dan manipulasi data jadwal, sekaligus mengadopsi kategori tugas (*task-archetype*) dari *benchmark* dan literatur terkini di bidang *question answering* (QA) dan *information retrieval* (IR). Detail dari klasifikasi skenario pertanyaan diberikan pada Tabel 1.

TABEL 1
KLASIFIKASI SKENARIO DAN REFERENSI KATEGORI TUGAS

Kode	Nama Kategori	Kategori Tugas/ *Benchmark*	Penjelasan	Ref.
K-1	*Lookup* Nama	*Single-span Factoid QA*	Mengambil atribut tunggal berbasis nama mahasiswa/ dosen	[17]
K-2	*Lookup* NIM	*Key-value Lookup*	Mengambil atribut tunggal berbasis NIM	[17]
K-3	Filter Multi-Atribut	Boolean/AND *Retrieval*	Filter banyak kriteria (misal ruang, tanggal, waktu)	[18]
K-4	*List-Retrieval*	*Small-set Recall* & *Precision*	Daftar entitas (misal daftar mahasiswa, dosen, NIM)	[19]
K-5	Agregasi Numerik	*Discrete Reasoning* / *Counting*	Agregasi jumlah jadwal/mahasiswa (COUNT)	[20]
K-6	Superlatif Deterministik	*Compositional Reasoning* (argmax/min)	Mencari entitas ekstrem (paling awal, paling akhir, dsb.)	[21]

Enam kategori skenario yang digunakan pada eksperimen ini disusun berdasarkan *task archetype* yang telah diadopsi secara luas dalam penelitian QA/IR, meliputi *lookup, multi-attribute filter, list retrieval, counting, dan compositional reasoning*. Pengelompokan ini bertujuan untuk merepresentasikan variasi tipe tugas ekstraksi yang relevan secara praktis dan teoretis, memfasilitasi perbandingan langsung dengan temuan riset sebelumnya,

memisahkan beban kognitif LLM antara format data terstruktur dan tidak terstruktur. Seluruh kategori di Tabel 1 telah disesuaikan dengan *task archetype* dari *benchmark* QA/IR populer, seperti SQuAD [17], DROP [20], serta penelitian terbaru terkait *retrieval-augmented* dan *compositional* QA [18], [19], [21]. Kategori ini juga dipilih agar beban pemrosesan LLM terhadap format data terstruktur dan tidak terstruktur dapat dibandingkan secara adil dan sistematis. Sebaran banyaknya pertanyaan di setiap kategori, beserta contoh pertanyaan diberikan pada Tabel 2. Setiap skenario dikodekan dalam fail metadata JSONL dengan atribut kunci: id, id_kategori, jenis kegiatan (seminar/sidang), semester (ganjil/genap), pekan, *prompt* (instruksi bahasa alami), dan *output_type* (jenis data yang diharapkan sebagai luaran). Selain daftar pertanyaan juga perlu dibuat jawaban (*ground-truth*) dari 200 pertanyaan. Proses pembuatan *ground-truth* menggunakan kueri SQL. Data yang tersimpan di fail *ground-truth* adalah data yang sesuai dengan jawaban skenario. Validasi dari fail *ground-truth* kemudian diperiksa kembali oleh validator (dua peneliti).

TABEL 2
CONTOH PERTANYAAN DAN JUMLAH SKENARIO PERTANYAAN DARI SETIAP KATEGORI

Kategori	Contoh Pertanyaan	Jumlah Skenario
K-1	Apa judul skripsi Ahmad Fikri Maulana?	80
K-2	Siapa nama mahasiswa dengan NIM 10120750?	14
K-3	Jadwal di R6010, 21 Agustus 2024, mulai 09.15	16
K-4	Daftar mahasiswa di ruang R6034 tanggal 19 Desember 2023	29
K-5	Berapa seminar di R6034 tanggal 19 Desember 2023?	22
K-6	Siapa mahasiswa dengan seminar paling pagi pada 25 Juni 2024?	39
	Total	200

Selain mengevaluasi performansi model LLM berdasarkan pertanyaan masukan, dilakukan pula evaluasi berdasarkan luaran yang dihasilkan dari pertanyaan masukan yang diberikan. Berdasarkan luarannya, penelitian ini mengelompokkan luarannya menjadi tiga kelompok yaitu: *schedule_rows, lecturer_list,* dan *aggregation*. *Schedule_rows* adalah bagaimana model LLM mampu menampilkan baris tabel jadwal secara lengkap atau dengan kata lain luaran dari semua entitas adalah relevan dengan jawaban. Untuk *lecturer_list* bertujuan untuk mengukur sejauh mana model LLM mampu memunculkan daftar nama dosen dengan lengkap dan benar. Yang terakhir adalah *aggregation*, parameter ini digunakan untuk menilai kemampuan LLM untuk dapat melakukan agregasi/perhitungan jumlah dari total seminar/sidang sesuai dengan pertanyaan yang diberikan. Detail sebaran data pertanyaan berdasarkan luarannya diberikan pada Tabel 3.

TABEL 3
SEBARAN JUMLAH DATA PERTANYAAN BERDASARKAN TIPE LUARAN DARI MODEL LLM

Tipe Luaran	Jumlah Skenario
schedule_rows	162
aggregation	22
lecturer_list	16
Total	200

B. Model LLM

Penelitian ini menguji tiga varian Gemini Flash: Gemini 2.5 Flash Preview (konteks 65.535 token, *knowledge cut-off* Januari 2025) serta Gemini 2.0 Flash dan Gemini 2.0 Flash-Lite (8.192 token, *cut-off* Juni 2024) [14], [15], [16]. Jumlah parameter ketiga model tidak dipublikasikan oleh Google AI. Seluruh model diakses melalui Google Generative AI SDK dengan *temperature* 0,0 (luaran deterministik) dan tanpa *fine-tuning*; performa dievaluasi secara *zero-shot* menggunakan *prompt* terstruktur.

C. Metrik Penilaian

Evaluasi performa model LLM dilakukan secara kuantitatif dengan menggunakan empat metrik utama yang umum digunakan pada ekstraksi entitas dan *question answering*:

1) Presisi (Precision): Presisi adalah metrik untuk mengukur proporsi prediksi benar dari seluruh hasil yang diprediksi model. Nilai presisi dihitung menggunakan persamaan (1).

$$\boldsymbol{Presisi} = \frac{TP}{(TP+FP)} \qquad (1)$$

2) Recall: Recall adalah metrik yang mengukur proporsi prediksi benar dari seluruh entitas yang seharusnya ditemukan model. Untuk nilai *recall* dihitung dengan menggunakan persamaan (2).

$$\boldsymbol{Recall} = \frac{TP}{(TP+FN)} \qquad (2)$$

Di mana:
True Positive (TP) adalah entitas yang diprediksi model dan benar ada di ground-truth.
False Positive (FP) adalah entitas yang diprediksi model tetapi tidak ada di ground-truth.
False Negative (FN) adalah entitas di ground-truth yang tidak berhasil diprediksi model.

3) F1-Score: Nilai F1 merupakan rata-rata harmonik dari presisi dan *recall* yang menunjukkan penilaian seimbang antara keduanya. Detail perhitungan nilai F1 ditunjukkan pada persamaan (3).

$$\boldsymbol{F1} = \frac{2\ x\ Presisi\ x\ Recall}{(Presisi+Recall)} \qquad (3)$$

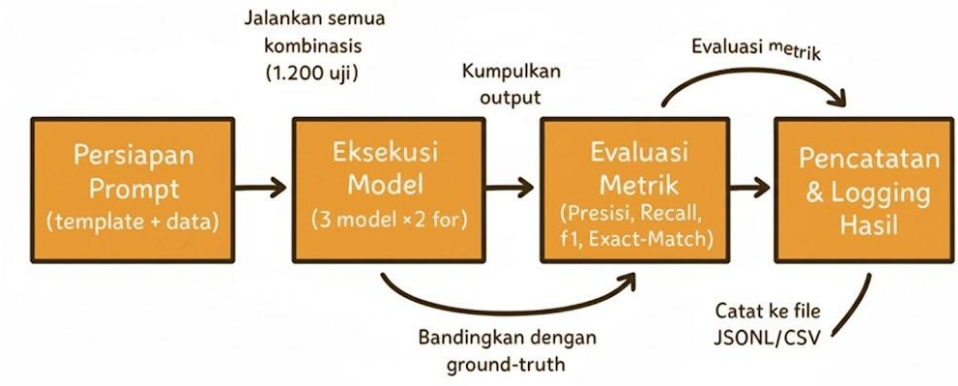
4) Exact-Match: Exact-Match adalah skor biner yang bernilai 1 jika luaran model identik sepenuhnya dengan *ground-truth*, dan 0 jika terdapat perbedaan apa pun. Untuk kategori berbasis *list* atau daftar, penilaian dilakukan secara *set equality* (tidak memperhatikan urutan). Pada entitas dengan kemungkinan variasi penulisan (judul skripsi, nama dosen), digunakan string *similarity* dengan nilai ambang 0,96; pada entitas eksak (NIM, ruang, waktu), digunakan *exact-match*.

Semua metrik dievaluasi pada level skenario, kemudian dirata-ratakan pada level kombinasi model dan format data. Selain evaluasi metrik, dilakukan uji signifikansi untuk menentukan apakah terdapat perbedaan performa antara format data JSON dan TXT. Pengujian menggunakan *paired t-test* dua ekor pada 200 pasang skor F1 dari skenario yang sama. Rumusan hipotesis:

$$\mathbf{H_0}: \mu_{JSON} = \mu_{TXT},\ \mathbf{H_1}: \mu_{JSON} \neq \mu_{TXT} \qquad (4)$$

Uji dilakukan pada tingkat signifikansi α=0,05.

D. Alur Pengujian



Gambar 3. Alur Pengujian Model LLM Gemini

Alur pengujian dalam penelitian ini dirancang secara sistematis sebagaimana ditunjukkan pada

Gambar 3, untuk memastikan setiap model LLM dievaluasi secara terstruktur dan konsisten. Diawali dengan memasukkan metadata skenario (pertanyaan), data *ground-truth* (jawaban), serta dua format konteks data (TXT dan JSON) yang relevan. *Prompt* pengujian disusun secara dinamis, menggabungkan instruksi bahasa alami dan data konteks, sesuai *template* untuk setiap kategori pertanyaan. *Prompt* dikirimkan ke *endpoint* API model Gemini yang telah ditentukan. Setiap kombinasi model (tiga varian), format data (TXT/JSON), dan skenario diuji secara terpisah, menghasilkan luaran model dalam format terstruktur (JSON) sesuai skema yang diharapkan. Latensi waktu eksekusi dan jumlah token (*prompt* + respons) dicatat otomatis untuk setiap panggilan model. Setiap entri luaran diekstraksi ke dalam struktur data yang seragam dengan ground-truth, guna menghindari bias akibat perbedaan format luaran model. Hasil luaran model dan *baseline* dievaluasi terhadap *ground-truth* menggunakan metrik akurasi (Presisi, *Recall*, F1, *Exact-Match*) secara otomatis untuk seluruh skenario. Evaluasi dilakukan untuk setiap pasangan format data (TXT vs JSON) pada model yang sama untuk memudahkan analisis perbedaan performa antar format. Seluruh hasil pengujian,

skor evaluasi, waktu eksekusi, dan detail lainnya dicatat secara sistematis ke fail JSONL dan CSV. Proses *logging* dilakukan *real-time* selama *pipeline* berjalan, menjamin seluruh rekam jejak pengujian terdokumentasi dengan baik dan dapat ditelusuri.

E. Analisis Kesalahan

Analisis kesalahan dilakukan untuk mengidentifikasi pola kegagalan yang tidak tercermin oleh metrik makro. Setiap prediksi dievaluasi pada tingkat skenario; prediksi dengan skor F1 = 0 ditandai sebagai kegagalan total. Jumlah kegagalan total dihitung per varian model, format konteks, dan kategori tugas (K-1 s.d. K-6). Selain itu, nilai *False Negative* (FN) dan *False Positive* (FP) diakumulasi untuk membangun peta panas distribusi kesalahan. Seluruh perhitungan dijalankan secara otomatis dengan skrip Python, dan hasilnya divisualisasikan dalam *heatmap* untuk memudahkan penafsiran pola kelemahan model pada tugas agregasi numerik serta superlatif deterministik.

III. HASIL DAN PEMBAHASAN

Bab ini menguraikan temuan dari serangkaian pengujian performa tiga varian model Gemini Flash pada tugas ekstraksi jadwal seminar dan sidang skripsi. Model diuji pada dua skenario masukan: teks PDF tidak terstruktur dan data JSON terstruktur. Kinerja dievaluasi menggunakan *precision*, *recall*, *F1-score*, dan *exact-match*. Selain membandingkan performa antar format, analisis menelaah sebaran skor pada enam kategori pertanyaan uji serta perbedaan latensi inferensi antar model. Seluruh pembahasan difokuskan pada evaluasi performa dasar tanpa membahas implementasi praktisnya.

A. Perbandingan Performansi Model Gemini

Pengujian pertama membandingkan performansi antar ketiga model Gemini baik dari performansi, latensi, jumlah token maupun efisiensi pada masing-masing data terstruktur (JSON) dan tidak terstruktur (TXT). Pada Tabel 4 ditunjukkan bahwa G2.5-Preview menjadi model yang paling baik dari keempat metrik yang digunakan. Untuk G2-Flash nilai presisi, *recall*, F1 dan nilai *exact-match* semuanya mencapai nilai yang sama baiknya, yaitu rata-rata dinilai 0,84 untuk data terstruktur dan untuk data tidak terstruktur mendapatkan nilai 0,01 lebih rendah. Hal ini menunjukkan bahwa hampir tidak ada perbedaan antara ekstraksi data terstruktur dan tidak terstruktur pada G2-Flash. Hasil yang berbeda diperoleh saat mengukur ketepatan hasil yang keluar yaitu sebesar 0,77 untuk data terstruktur dan 0,78 untuk data tidak terstruktur. Hal ini disebabkan karena luaran dari model G2-Flash tidak lengkap atau terpotong sehingga pada pengukuran *exact-match* hasil luaran dianggap gagal memunculkan hasil yang benar. Temuan ini juga diperkuat dari penelitian [10] yang menyatakan bahwa LLM masih mungkin mengalami kesulitan dalam memahami struktur internal dan komposisi token walaupun telah model telah diadaptasi dengan *one-shot* ataupun *few-shot learning*. Untuk G2-Lite pada data tidak terstruktur, jawaban yang diharapkan terpanggil mencapai 0,81 namun di antara jawaban yang dimunculkan oleh model G2-Lite ketepatan jawaban tersebut hanya sebesar 0,79. Hasil ini juga diikuti dengan persentase *exact-match* yang menurun mencapai 0,68 jika dibandingkan dengan model G2-Flash. Temuan ini menguatkan ulasan penelitian sebelumnya yang menyatakan bahwa selain kompleksitas tugas yang harus dikerjakan oleh model, banyaknya parameter menentukan performansi dari model tersebut [11].

TABEL 4
PERFORMA GEMINI FLASH TERHADAP DATA TERSTRUKTUR DAN TIDAK TERSTRUKTUR

Model	Sumber	Presisi	Recall	F1	Exact-Match
G2.5-Preview	*json_source*	0,9888	0,9900	0,9893	0,9850
	txt_source	0,9850	0,9825	0,9833	0,9800
G2-Flash	*json_source*	0,8541	0,8457	0,8441	0,7750
	txt_source	0,8386	0,8398	0,8351	0,7800
G2-Lite	*json_source*	0,8018	0,8161	0,7982	0,7200
	txt_source	0,7911	0,8109	0,7899	0,6850

B. Evaluasi Model Berdasarkan Tipe Luaran

Tiga tipe luaran (*schedule_rows*, *lecturer_list*, dan *aggregation*) telah didefinisikan pada Bagian II-A. Tabel 5 merangkum nilai F1 ketiga varian Gemini Flash pada masing-masing tipe luaran untuk konteks JSON dan TXT. Secara umum, Gemini 2.5 Flash Preview mempertahankan kinerja tertinggi pada seluruh tipe luaran (F1 ≥ 0,98). Pada luaran baris jadwal (*schedule_rows*) dan daftar dosen (*lecturer_list*), Gemini 2.0 Flash konsisten melampaui Gemini 2.0 Flash-Lite, kecuali pada daftar dosen berbasis TXT, di mana selisih keduanya tipis. Performa agregasi masih menjadi tantangan bagi dua model berkapasitas lebih kecil: F1 ≈ 0,41 pada JSON dan tetap rendah pada TXT. Hal ini menunjukkan bahwa operasi perhitungan numerik memerlukan konteks yang lebih panjang atau kapasitas parameter yang lebih besar. Dibandingkan format data TXT, format JSON cenderung meningkatkan F1 model berkapasitas

menengah-ringan (G2-Flash, G2-Lite) khususnya pada daftar dosen dan baris jadwal, tetapi tidak cukup menaikkan kinerja agregasi. Dengan demikian, struktur data terformat bermanfaat bagi model kecil–menengah, sedangkan model besar andal bahkan tanpa prapemrosesan tambahan.

TABEL 5
PERBANDINGAN NILAI F1 DARI KETIGA MODEL GEMINI FLASH

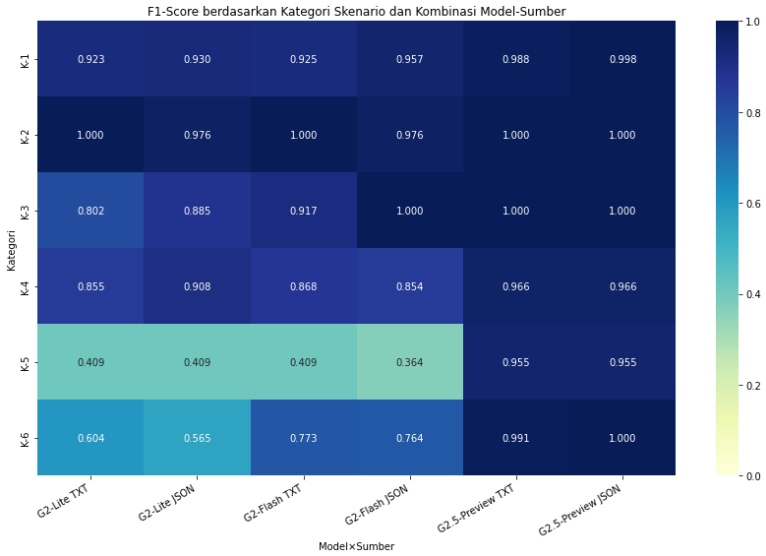
Tipe Luaran	JSON			TXT		
	G2-Lite	G2-Flash	G2.5-Preview	G2-Lite	G2-Flash	G2.5-Preview
aggregation	0,4090	0,3640	0,9550	0,4090	0,4090	0,9550
lecturer_list	0,8570	0,8620	1,0000	0,8830	0,7750	0,9790
schedule_rows	0,8450	0,9080	0,9930	0,8320	0,8990	0,9880

C. Evaluasi Model Berdasarkan Kategori Pertanyaan

Berdasarkan kategori pertanyaan yang diajukan maka pertanyaan K-1 dan K-2, yaitu pertanyaan yang menggunakan atribut tunggal (NIM, Nama) mampu dijawab dengan baik oleh ketiga model Gemini. Untuk kategori pertanyaan K-3, yang menggunakan multiatribut, mampu diidentifikasi oleh G2-Flash dan G2.5-Preview dengan baik, terutama untuk data terstruktur dengan format JSON. Pada Tabel 6, model G2.5-Preview mampu bekerja sama baiknya menggunakan data terstruktur maupun tidak terstruktur di semua kategori pertanyaan. Data dengan format JSON untuk pertanyaan kategori K-3 dan K-4 menghasilkan nilai F1 yang lebih baik saat menggunakan model G2-Lite dan G2-Flash. Pertanyaan K-5 tentang agregasi numerik masih menjadi tantangan bagi model G2-Lite dan G2-Flash. Penggunaan data terstruktur dengan format JSON ataupun tidak terstruktur dengan format TXT tidak meningkatkan nilai F1 dari kedua model tersebut. Pertanyaan K-6 yaitu tentang superlatif deterministik dapat menghasilkan nilai F1 yang lebih tinggi ketika menggunakan data TXT pada model G2-Lite dan G2-Flash. Walaupun begitu, nilai F1 dari kedua model ini tidak lebih baik dari G2.5-Preview.

TABEL 6
NILAI F1 TIGA VARIAN GEMINI FLASH PER KATEGORI PERTANYAAN DAN TIPE DATA

Kategori	JSON			TXT		
	G2-Lite	G2-Flash	G2.5-Preview	G2-Lite	G2-Flash	G2.5-Preview
K-1	0,9300	0,9570	0,9980	0,9230	0,9250	0,9880
K-2	0,9760	0,9760	1,0000	1,0000	1,0000	1,0000
K-3	0,8850	1,0000	1,0000	0,8020	0,9170	1,0000
K-4	0,9080	0,8540	0,9660	0,8550	0,8680	0,9660
K-5	0,4090	0,3640	0,9550	0,4090	0,4090	0,9550
K-6	0,5650	0,7640	1,0000	0,6040	0,7730	0,9910

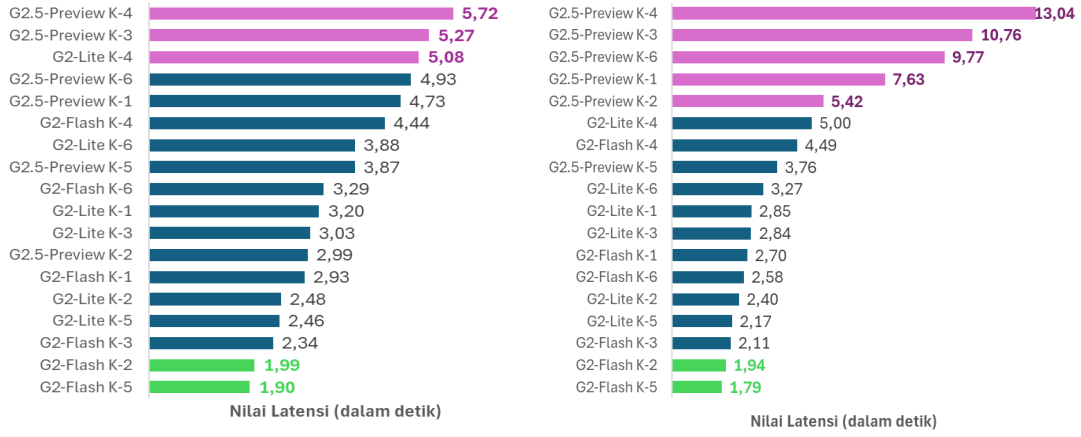


Gambar 4. *Heatmap* Skor F1 Gemini Flash (TXT vs JSON)

Gambar 4 memperlihatkan pola distribusi skor F1 untuk setiap kategori pertanyaan pada ketiga varian Gemini Flash dan dua format data masukan. Terlihat jelas pada *heatmap* bahwa kinerja G2.5-Preview konsisten tinggi (≥ 0,95) di hampir seluruh kategori dan format, sedangkan G2-Lite menunjukkan penurunan kontras khususnya pada kategori agregasi (K-5) dan superlatif deterministik (K-6). Warna yang lebih terang pada baris K-5 milik G2-Flash menegaskan kembali kesulitan model ringan dalam menangani operasi numerik, sedangkan area warna pekat milik G2.5-Preview memperlihatkan stabilitas performa lintas kategori. Perbedaan warna antara sel JSON dan TXT di baris K-3 / K-4 pada G2-Flash mengonfirmasi keuntungan data terstruktur yang telah dibahas sebelumnya, namun efek tersebut menghilang pada model G2.5-Preview. Hal ini mengindikasikan kapasitas model lebih besar menetralkan ketergantungan terhadap format konteks.

D. Evaluasi Latensi

Rata-rata waktu tanggap (latensi) menjadi pertimbangan penting dalam penerapan model LLM untuk aplikasi praktis. Hasil pengujian menunjukkan bahwa rata-rata latensi per permintaan adalah 1,8 detik pada Gemini 2.0 Flash-Lite, 2,0 detik pada Gemini 2.0 Flash, dan 3,9 detik pada Gemini 2.5 Flash Preview.



(a) (b)

Gambar 5. Rata-Rata Latensi Gemini Flash per Kategori: (a) JSON, (b) TXT

Secara umum, perbedaan latensi antar kategori pertanyaan (K-1 hingga K-6) tidak terlalu signifikan, dan pola latensi konsisten pada seluruh kategori, sebagaimana dapat dilihat pada Gambar 5(a) untuk format JSON dan Gambar 5(b) untuk format TXT. Format JSON secara konsisten menghasilkan latensi yang sedikit lebih rendah dibandingkan format TXT, dengan rata-rata selisih sekitar 0,2 detik. Hal ini diduga karena prapemrosesan data lebih kompleks pada format teks mentah. Walaupun Gemini 2.5 Flash Preview memiliki latensi tertinggi, varian ini tetap menunjukkan akurasi terbaik (lihat Bagian III-A), sehingga menegaskan *trade-off* antara kecepatan dan kapasitas model. Seluruh respons yang dihasilkan pada pengujian ini memiliki panjang ≤ 700 token, sehingga panjang respons tidak menjadi pembatas performa. Log latensi dan token lengkap tersedia di repositori GitHub [22].

E. Uji Signifikansi Format Data (Paired t-Test)

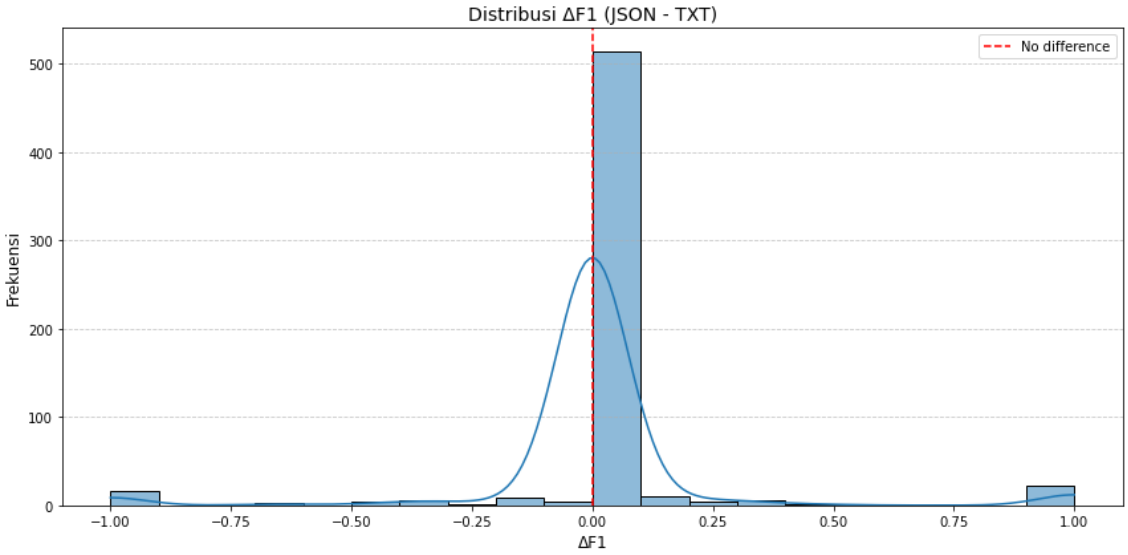
Signifikansi perbedaan skor F1 antara JSON dan TXT diuji dengan *paired t-test* pada 600 pasang data. Hasil uji normalitas Shapiro–Wilk (p = 0,21) memenuhi asumsi normalitas. Statistik ringkasan ditampilkan Tabel 7.

TABEL 7
RINGKASAN STATISTIK UJI *PAIRED T* ΔF1 (JSON VERSUS TXT)

Statistik	Nilai
ΔF1	0,0077
SD(ΔF1)	0,2651
t(599)	0,7141
p-value	0,4755
Cohen's d_{paired}	0,03
95 % CI	[–0,014; 0,030]

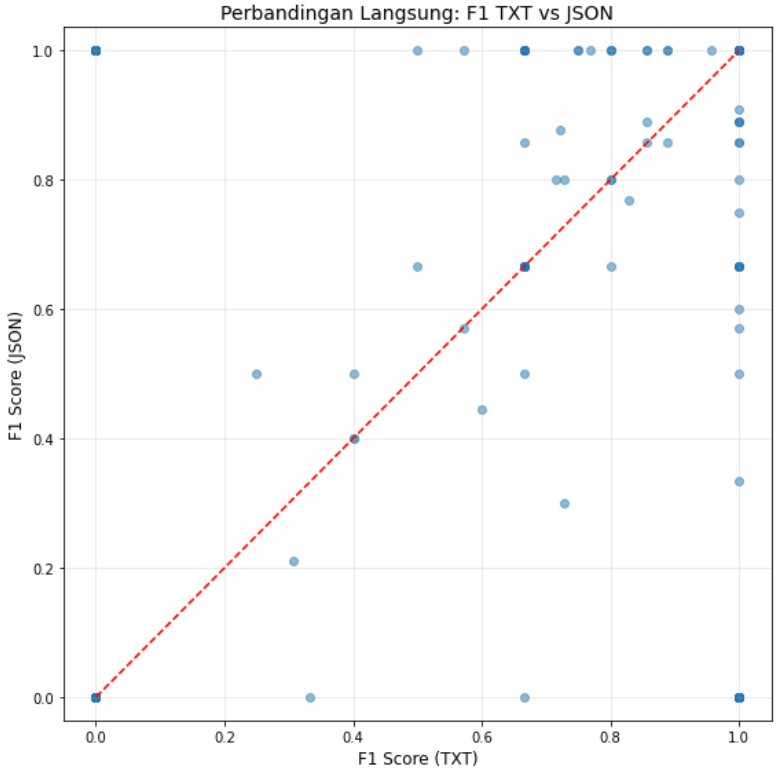
Nilai p-value sebesar 0,48 (p > 0,05) mengindikasikan bahwa hipotesis nol (H_0 : μ_JSON = μ_TXT) gagal ditolak, artinya tidak terdapat bukti statistik yang cukup untuk menyatakan adanya perbedaan performa. Sejalan

dengan anjuran praktik statistik modern yang menekankan pentingnya melampaui p-value [23], analisis ukuran efek menunjukkan nilai Cohen's d = 0,03. Ukuran ini tergolong sangat kecil jika dibandingkan dengan tolok ukur konvensional untuk efek kecil (d ≈ 0,20) [24], sehingga dapat disimpulkan bahwa perbedaan yang teramati tidak signifikan baik secara statistik maupun praktis.



Gambar 6. Histogram Selisih Skor F1 (ΔF1 = F1JSON − F1TXT, n = 600)

Histogram selisih skor ΔF1 yang terpusat di sekitar nol menunjukkan bahwa rata-rata perbedaan antara skor F1 pada konteks JSON dan TXT mendekati nol, sehingga tidak terdapat kecenderungan sistematis ke salah satu format. Meski demikian, interpretasi hasil tersebut perlu mempertimbangkan pula lebar sebaran dan batas kesepakatan, karena distribusi yang terpusat saja tidak menjamin kesetaraan kinerja dua metode. Pendekatan ini sejalan dengan panduan terkini mengenai evaluasi kesepakatan antar-metode yang menekankan pentingnya visualisasi selisih disertai ukuran efek dan interval kepercayaan untuk memperoleh kesimpulan yang lebih informatif [25].



Gambar 7. Sebaran Skor F1 TXT versus JSON dengan Garis Identitas (n = 600)

Scatter plot pada Gambar 7 memperlihatkan sebagian besar titik data berada sangat dekat dengan garis identitas, menandakan bahwa nilai F1 antara masukan JSON dan TXT hampir sama pada seluruh pengamatan. Pola tersebut mengindikasikan tidak adanya bias sistematis yang menguntungkan salah satu format, karena penyimpangan terhadap garis identitas hanya terjadi secara acak dan berskala kecil. Visualisasi ini sekaligus memperkuat hasil uji statistik dan ukuran efek sebelumnya, yang menunjukkan bahwa kinerja ketiga model Gemini 2.5 Flash konsisten di antara kedua jenis masukan.

F. Analisis Kesalahan Ringkas

Analisis kesalahan secara ringkas dilakukan dengan mengevaluasi pola kegagalan ekstraksi berdasarkan jumlah kasus dengan skor F1=0 (kegagalan total) dan distribusinya terhadap format konteks serta kategori tugas yang diuji. Ringkasan kegagalan total dan tingkat *Exact-Match* untuk setiap varian model pada dua format konteks disajikan dalam Tabel 8. Varian Gemini 2.0 Flash-Lite memperlihatkan jumlah kegagalan total terbesar, yaitu sebanyak 65 kasus (53%) dari total 122 kasus, sedangkan varian Gemini 2.5 Flash Preview paling sedikit mengalami kegagalan dengan hanya lima kasus (4%) dan tingkat *Exact-Match* sebesar 98%.

TABEL 8
RINGKASAN ANALISIS KESALAHAN PER MODEL DAN FORMAT KONTEKS

Model	Konteks	#ID F1=0	Exact-Match (%)
Gemini 2.0 Flash-Lite	TXT	33	68 (137/200)
Gemini 2.0 Flash-Lite	JSON	32	72 (144/200)
Gemini 2.0 Flash	TXT	27	78 (156/200)
Gemini 2.0 Flash	JSON	25	78 (155/200)
Gemini 2.5 Flash Preview	TXT	3	98 (196/200)
Gemini 2.5 Flash Preview	JSON	2	98 (197/200)

Selanjutnya, untuk mendapatkan gambaran yang lebih detail mengenai tipe tugas yang paling sering menyebabkan kegagalan fatal, dilakukan analisis distribusi kasus dengan skor F1=0 berdasarkan kategori tugas (K-1 hingga K-6). Ringkasan data agregat pada Tabel 9 memperlihatkan bahwa sebagian besar kegagalan fatal terjadi pada kategori yang memerlukan pemrosesan informasi kompleks, yaitu kategori agregasi numerik (K-5) sebesar 45% (55 kasus) dan superlatif deterministik (K-6) sebesar 33% (40 kasus). Dengan demikian, kedua kategori ini menyumbang sekitar 78% dari total kegagalan fatal, sementara kategori yang lebih sederhana seperti *lookup* tunggal (K-1 dan K-2) hanya berkontribusi sedikit atau bahkan tidak menyumbang kegagalan sama sekali.

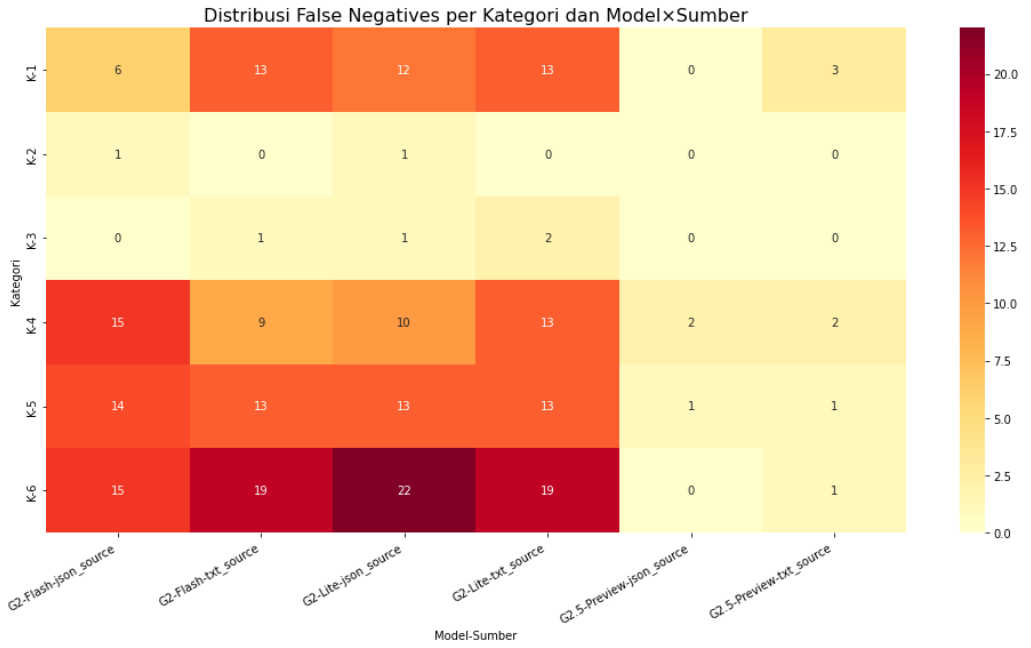
TABEL 9
DISTRIBUSI KEGAGALAN TOTAL (F1=0) PER KATEGORI TUGAS (AGREGAT SEMUA MODEL)

Kategori	K-1	K-2	K-3	K-4	K-5	K-6	Total
#ID F1=0	15	0	4	8	55	40	122
Proporsi	12%	0%	3%	7%	45%	33%	100%

Distribusi kesalahan yang lebih spesifik untuk tiap varian model di dua kategori paling rentan tersebut ditampilkan dalam Tabel 10. Varian Gemini 2.0 Flash-Lite memiliki jumlah kegagalan tertinggi baik di K-5 maupun K-6, masing-masing dengan proporsi kegagalan sebesar 59% dan 32%, dibandingkan dengan varian Gemini 2.0 Flash dan Gemini 2.5 Flash Preview yang menunjukkan angka kegagalan yang jauh lebih rendah.

TABEL 10
DISTRIBUSI KEGAGALAN FATAL (F1=0) PER KATEGORI TUGAS PADA TIAP MODEL

Model	K-5 (Agregasi Numerik)	K-6 (Superlatif Deterministik)
Gemini 2.0 Flash-Lite	26 dari 44 (59%)	25 dari 78 (32%)
Gemini 2.0 Flash	27 dari 44 (61%)	15 dari 78 (19%)
Gemini 2.5 Flash Preview	2 dari 44 (4.5%)	1 dari 78 (1.3%)



Gambar 8. *Heatmap* Distribusi False Negatives (FN) per Kategori Tugas dan Model

Untuk memperjelas pola kesalahan secara visual, Gambar 8 menyajikan peta panas (*heatmap*) distribusi jumlah kasus *False Negative* (FN) untuk tiap kategori tugas terhadap ketiga varian model dengan kedua format konteks yang diuji. Berdasarkan visualisasi tersebut, intensitas FN tertinggi terjadi pada kategori K-5 (agregasi numerik) dan K-6 (superlatif deterministik), terutama untuk varian Gemini 2.0 Flash-Lite, yang tampak melalui warna lebih gelap pada kedua kategori tersebut. Sebaliknya, tugas yang lebih sederhana seperti lookup tunggal (K-1 dan K-2) menunjukkan performa lebih stabil dengan tingkat *Exact-Match* di atas 88 % di seluruh model. Secara keseluruhan, hasil ini menunjukkan bahwa tugas yang melibatkan pemrosesan angka dan penentuan nilai ekstrem masih menjadi tantangan utama bagi model bahasa berkapasitas kecil; karena itu, implementasi nyata disarankan menambahkan pendekatan tambahan seperti *prompt* numerik eksplisit atau pascaproses berbasis aturan agar keterbacaan hasil tetap konsisten dan akurat.

IV. SIMPULAN

Penelitian ini menunjukkan bahwa seluruh varian Gemini mampu mengekstraksi entitas jadwal seminar dan sidang skripsi secara andal, dengan kinerja yang sebanding antara format masukan teks tidak terstruktur dan JSON terstruktur. Performa model sangat dipengaruhi oleh kapasitasnya, di mana varian berkapasitas tinggi menghasilkan presisi dan stabilitas yang lebih baik dibanding model menengah dan ringan. Tantangan utama masih ditemukan pada tugas yang melibatkan penalaran numerik kompleks, sedangkan data yang lebih terstruktur membantu mengurangi kesalahan pada model ringan. Temuan ini menegaskan pentingnya pemilihan model dan format data yang sesuai untuk menjaga akurasi dan efisiensi pada skenario ekstraksi informasi berbasis LLM.

UCAPAN TERIMAKASIH

Penelitian ini dibiayai oleh Universitas Komputer Indonesia (UNIKOM) melalui fasilitasi Direktorat Penelitian, Pengabdian dan Pemberdayaan Masyarakat (DP3M), berdasarkan Kontrak Pelaksanaan Penelitian Nomor: 030/SP/DP3M/UNIKOM/XII/2024. Penulis mengucapkan terima kasih atas dukungan finansial dan fasilitas yang diberikan oleh UNIKOM. Apresiasi juga disampaikan kepada Fadel Anugerah Gusti dan Stefanus Gratilio, yang turut mendukung proses dengan memberikan masukan awal sebagai tester prototipe aplikasi, serta kepada semua pihak yang turut berkontribusi dalam penyelesaian penelitian ini.

DAFTAR PUSTAKA

[1] M. Arnold, M. Goldschmitt, and T. Rigotti, "Dealing with information overload: a comprehensive review," *Front. Psychol.*, vol. 14, June 2023, doi: 10.3389/fpsyg.2023.1122200.
[2] S. Al Banna, I. Rakhmani, and D. C. Sukmajati, "Kilas Kebijakan PSPK – Membangun Keahlian dengan Profesionalisme di Kampus Merdeka: Kajian mengenai Beban Kerja Dosen di Indonesia," Pusat Studi Pendidikan dan Kebijakan (PSPK), Jakarta, Apr. 2022. [Online]. Available: https://pspk.id/wp-content/uploads/2022/04/Kilas-Kebijakan-PSPK-Membangun-Keahlian-dengan-Profesionalisme-di-Kampus-Merdeka-Kajian-Mengenai-Beban-Kerja-Dosen-di-Indonesia.pdf
[3] OECD, "Ensuring quality in VET and higher education: Getting quality assurance right," OECD Education Policy Perspectives, Feb. 2025. doi: 10.1787/812ff006-en.
[4] F. P. Diallo and C. Tudose, "Optimizing the Scheduling of Teaching Activities in a Faculty," *Appl. Sci.*, vol. 14, no. 20, Art. no. 20, Jan. 2024, doi: 10.3390/app14209554.
[5] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf
[6] D. Xu *et al.*, "Large language models for generative information extraction: a survey," *Front. Comput. Sci.*, vol. 18, no. 6, p. 186357, Nov. 2024, doi: 10.1007/s11704-024-40555-y.
[7] B. Grothey *et al.*, "Comprehensive testing of large language models for extraction of structured data in pathology," *Commun. Med.*, vol. 5, no. 1, pp. 1–13, Mar. 2025, doi: 10.1038/s43856-025-00808-8.
[8] V. Ntinopoulos *et al.*, "Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation," *BMJ Health Care Inform.*, vol. 32, no. 1, p. e101139, Jan. 2025, doi: 10.1136/bmjhci-2024-101139.
[9] A. Meddeb *et al.*, "Evaluating local open-source large language models for data extraction from unstructured reports on mechanical thrombectomy in patients with ischemic stroke," *J. NeuroInterventional Surg.*, Aug. 2024, doi: 10.1136/jnis-2024-022078.
[10] D. Jobson and Y. Li, "Investigating the Potential of Using Large Language Models for Scheduling," in *Proceedings of the 1st ACM International Conference on AI-Powered Software*, in AIware 2024. New York, NY, USA: Association for Computing Machinery, July 2024, pp. 170–171. doi: 10.1145/3664646.3665084.
[11] Y. Chai, Y. Fang, Q. Peng, and X. Li, "Tokenization Falling Short: On Subword Robustness in Large Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1582–1599. doi: 10.18653/v1/2024.findings-emnlp.86.
[12] T. A. Chang and B. K. Bergen, "Language Model Behavior: A Comprehensive Survey," *Comput. Linguist.*, vol. 50, no. 1, pp. 293–350, Mar. 2024, doi: 10.1162/coli_a_00492.
[13] D. Balsiger, H.-R. Dimmler, S. Egger-Horstmann, and T. Hanne, "Assessing Large Language Models Used for Extracting Table Information from Annual Financial Reports," *Computers*, vol. 13, no. 10, Art. no. 10, Oct. 2024, doi: 10.3390/computers13100257.
[14] Gemini Team and Google, "Gemini: A Family of Highly Capable Multimodal Models," Google, 2024. Accessed: May 28, 2025. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf
[15] "Gemini 2.5: Our most intelligent models are getting even better," Google. Accessed: June 09, 2025. [Online]. Available: https://blog.google/technology/google-deepmind/google-gemini-updates-io-2025/
[16] V. N. Box Head of Product Marketing, Platform at, "Gemini 2.5 Flash Delivers Enhanced Document Q&A and Extraction with Box AI," Box Blog. Accessed: June 09, 2025. [Online]. Available: https://backend.blog.box.com/gemini-25-flash-delivers-enhanced-document-qa-and-extraction-box-ai
[17] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds., Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. doi: 10.18653/v1/D16-1264.
[18] S. Krishna *et al.*, "Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds., Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 4745–4759. Accessed: May 04, 2025. [Online]. Available: https://aclanthology.org/2025.naacl-long.243/
[19] N. F. Liu *et al.*, "Lost in the Middle: How Language Models Use Long Contexts," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 157–173, 2024, doi: 10.1162/tacl_a_00638.
[20] D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner, "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 2368–2378. doi: 10.18653/v1/N19-1246.
[21] S. Neelam *et al.*, "SYGMA: A System for Generalizable and Modular Question Answering Over Knowledge Bases," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3866–3879. doi: 10.18653/v1/2022.findings-emnlp.284.
[22] G. Hermawan and E. Rainarli, "Evaluasi-LLM-Gemini-Jadwal-Skripsi — Dataset & Code." GitHub repository, June 02, 2025. Accessed: June 10, 2025. [Online]. Available: https://github.com/Galih-Hermawan-Unikom/evaluasi-llm-gemini-jadwal-skripsi
[23] L. N. Yaddanapudi, "The American Statistical Association statement on P-values explained," *J. Anaesthesiol. Clin. Pharmacol.*, vol. 32, no. 4, pp. 421–423, 2016, doi: 10.4103/0970-9185.194772.
[24] D. Lakens, "Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs," *Front. Psychol.*, vol. 4, p. 863, Nov. 2013, doi: 10.3389/fpsyg.2013.00863.
[25] P. S. P. Silveira, J. E. Vieira, and J. de O. Siqueira, "Is the Bland-Altman plot method useful without inferences for accuracy, precision, and agreement?," *Rev. Saúde Pública*, vol. 58, p. 01, doi: 10.11606/s1518-8787.2024058005430.