

# ANALISIS PERFORMA *LARGE LANGUAGE MODEL* DALAM PENILAIAN OTOMATIS MENGGUNAKAN MODEL CAMPURAN LINEAR

DALILAH HUSNA



PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA  
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2026

@Hak cipta milik IPB University

IPB University



IPB University  
Bogor Indonesia

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
    - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
    - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
  2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Perpustakaan IPB University



### @Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

## PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Analisis Performa *Large Language Model* dalam Penilaian Otomatis Menggunakan Model Campuran Linear” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dalam penyusunan karya ini penulis menggunakan bantuan kecerdasan buatan, yaitu Gemini 3.5 Flash untuk penyelarasan kalimat dan peninjauan paragraf, serta GPT 5.2 untuk kompilasi dan perbaikan kode analisis. Setelah menggunakan layanan tersebut, penulis meninjau dan menyunting konten sesuai kebutuhan serta bertanggung jawab penuh atas isi karya tugas akhir ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juni 2026

Dalilah Husna  
G1401221003



## ABSTRAK

DALILAH HUSNA. Analisis Performa *Large Language Model* dalam Penilaian Otomatis Menggunakan Model Campuran Linear. Dibimbing oleh BAGUS SARTONO dan CICI SUHAENI.

Penilaian manual pada soal isian membutuhkan waktu yang relatif lama dan berpotensi menimbulkan inkonsistensi antarpenilai. Penelitian ini bertujuan mengkaji performa tiga *Large Language Model* (LLM), yaitu Gemini, GPT, dan Claude, dalam melakukan penilaian otomatis (*autograding*) pada dua konteks, *Automatic Essay Scoring* (AES) dan *Automatic Short Answer Grading* (ASAG). Penilaian dilakukan menggunakan delapan skema *prompt* yang dirancang berdasarkan variasi pola instruksi, urutan output, serta keberadaan contoh. Konsistensi internal antara skor dan alasan penilaian dievaluasi menggunakan uji Mantel dan divisualisasikan melalui *Non-metric Multidimensional Scaling* (NMDS). Selanjutnya, perbedaan performa antarkombinasi model dan skema *prompt* dianalisis menggunakan model campuran linear dengan peubah respons berupa kuadrat selisih skor, yang mengakomodasi struktur hierarki data melalui penambahan efek acak. Hasil eksplorasi menunjukkan bahwa LLM cenderung lebih optimal dalam melakukan penilaian pada konteks ASAG. Penggunaan rubrik terbukti meningkatkan akurasi penilaian dibandingkan tanpa rubrik. Berdasarkan pengujian konsistensi, GPT menunjukkan keselarasan tertinggi antara skor dan alasan penilaian pada konteks AES, sedangkan Gemini unggul pada konteks ASAG. Analisis lebih lanjut menunjukkan adanya interaksi antara LLM dan skema *prompt* terhadap kinerja penilaian otomatis. Pada konteks AES, skema *prompt* yang melibatkan pemberian contoh menghasilkan performa yang lebih baik. Sementara itu, pada konteks ASAG, Gemini dan GPT menunjukkan performa optimal tanpa perbedaan signifikan.

Kata kunci : *large language model*, model campuran linear, *non-metric multidimensional scaling*, penilaian otomatis.

@Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

## ABSTRACT

DALILAH HUSNA. Performance Analysis of Large Language Model in Automatic Assessment Using Linear Mixed Model. Supervised by BAGUS SARTONO and CICI SUHAENI.

Manual grading of open-ended responses requires considerable time and may lead to inter-rater inconsistency. This study aims to examine the performance of three Large Language Models (LLMs), Gemini, GPT, and Claude, in automated scoring across two contexts, Automatic Essay Scoring (AES) and Automatic Short Answer Grading (ASAG). The evaluation employs eight prompt schemes designed based on variations in instruction patterns, output order, and the inclusion of examples. Internal consistency between assigned scores and their justifications is assessed using the Mantel test and visualized through Non-metric Multidimensional Scaling (NMDS). Furthermore, differences in performance across model prompt combinations are analyzed using a linear mixed-effects model, with the squared score difference as the response variable, accounting for the hierarchical data structure through random effects. The findings indicate that LLMs tend to perform better in the ASAG context. The use of scoring rubrics significantly improves grading accuracy compared to scenarios without rubrics. In terms of consistency, GPT demonstrates the highest alignment between scores and explanations in AES, while Gemini performs best in ASAG. Additionally, results reveal an interaction effect between LLM type and prompt scheme on grading performance. In AES, prompts incorporating examples yield better results, whereas in ASAG, Gemini and GPT achieve optimal performance without significant differences.

**Keywords:** autograding, large language model, linear mixed model, non-metric multidimensional scaling.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
    - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
    - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
  2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2026<sup>1</sup>  
Hak Cipta dilindungi Undang-Undang

*Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.*

*Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.*

# **ANALISIS PERFORMA *LARGE LANGUAGE MODEL* DALAM PENILAIAN OTOMATIS MENGGUNAKAN MODEL CAMPURAN LINEAR**

**DALILAH HUSNA**

Skripsi  
sebagai salah satu syarat untuk memperoleh gelar  
Sarjana pada  
Program Studi Statistika dan Sains Data

**PROGRAM STUDI SARJANA STATISTIKA DAN SAINS DATA  
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA  
INSTITUT PERTANIAN BOGOR  
BOGOR  
2026**



*@Hak cipta milik IPB University*

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Skripsi:  
Dr. Aam Alamudi, M.Si



### @Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
  - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
  - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Judul Skripsi : Analisis Performa *Large Language Model* dalam Penilaian Otomatis Menggunakan Model Campuran Linear

Nama : Dalilah Husna  
NIM : G1401221003

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:  
Dr. Bagus Sartono, S.Si., M.Si.

---

Pembimbing 2:  
Cici Suhaeni, S.Si., M.Si., Ph.D

---

Diketahui oleh

Ketua Program Studi:  
Dr. Bagus Sartono, S.Si., M.Si.  
NIP 197804112005011002

---

Tanggal Ujian:  
8 Juni 2026

Tanggal Lulus:

## PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Judul yang dipilih dalam penelitian ini “Analisis Performa *Large Language Model* dalam Penilaian Otomatis Menggunakan Model Campuran Linear”. Pada kesempatan ini, penulis ingin mengucapkan terima kasih kepada seluruh pihak yang telah berperan, sejak penulis menempuh studi sampai proses akhir penulisan karya ilmiah ini, di antaranya :

1. Bapak Suryadi dan Ibu Lila Widiyastuti serta Afif Fakhri Muhammad selaku keluarga penulis yang selalu memberikan doa, dukungan dan kasih sayang kepada penulis;
2. Bapak Dr. Bagus Sartono, S.Si., M.Si. dan Ibu Cici Suhaeni, S.Si., M.Si., Ph.D yang telah membimbing, memberi arahan dan saran dalam penulisan karya ilmiah ini;
3. Ibu Laily Nissa Atul Mualifah, M.Si. dan Ibu Sachnaz Desta Oktarina, S.Stat., M.Agr., Ph.D selaku dosen moderator kolokium dan seminar hasil yang telah memberikan saran dan masukan dalam penulisan karya ilmiah ini;
4. Bapak Ir. Aam Alamudi, M.Si. selaku dosen penguji pada sidang skripsi yang telah memberikan saran dan masukan pada penulisan karya ilmiah ini;
5. Seluruh dosen dan tenaga pendidik Program Studi Statistika dan Sains Data yang telah memberi ilmu dan menunjang segala kebutuhan penulis selama perkuliahan dan penyusunan karya ilmiah ini;
6. Ulfah, Nita, Cindy, Lilis, Azkiya, Biki dan Nafisa selaku teman penulis yang telah menemani dan memberi semangat selama penyusunan karya ilmiah ini;
7. Teman-teman Statistika 59 dan kepada seluruh pihak yang tidak dapat disebutkan satu per satu yang telah mendukung selama proses perkuliahan serta penyusunan karya ilmiah ini.

Penulis menyadari bahwa karya ilmiah ini masih belum mencapai tingkat kesempurnaan yang diharapkan. Oleh karena itu, penulis memohon maaf apabila ditemukan kesalahan dan kekurangan dalam karya ilmiah ini. Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juni 2026

*Dalilah Husna*



## DAFTAR ISI

DAFTAR TABEL	iii
DAFTAR GAMBAR	iii
DAFTAR LAMPIRAN	iii
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan	2
TINJAUAN PUSTAKA	3
2.1 Penilaian Soal Ujian	3
2.2 <i>Large Language Model</i>	3
2.3 <i>Prompt Engineering</i>	5
2.4 Penilaian Otomatis	5
2.5 Korelasi Spearman	7
2.6 <i>Non-metric Multidimensional Scaling</i>	7
2.7 Model Campuran Linear	8
2.8 Metode Rasio <i>Likelihood</i>	9
2.9 Uji Wald	10
2.10 <i>Estimated Marginal Means</i>	10
2.11 <i>Compact Letter Display</i>	11
III METODE	12
3.1 Data	12
3.2 Prosedur Analisis Data	12
IV HASIL DAN PEMBAHASAN	16
4.1 Skema Penilaian Otomatis	16
4.2 Eksplorasi Hasil Penilaian Otomatis	17
4.3 Evaluasi Konsistensi Output Penilaian Otomatis	22
4.4 Evaluasi Kesesuaian Skor Penilaian Otomatis dengan Dosen	27
V SIMPULAN DAN SARAN	32
5.1 Simpulan	32
5.2 Saran	32
DAFTAR PUSTAKA	33
LAMPIRAN	36
RIWAYAT HIDUP	38

## DAFTAR TABEL

1	Interpretasi koefisien korelasi Spearman	7
2	Keterangan data yang digunakan dalam penelitian	12
3	Skema <i>prompt</i> yang digunakan dalam penilaian otomatis	16
4	Nilai koefisien korelasi Spearman antara skor LLM dan dosen	18
5	Hasil uji Mantel untuk evaluasi kekonsistenan data AES	23
6	Hasil uji Mantel untuk evaluasi kekonsistenan data ASAG	25
7	Hasil perbandingan model dengan skema penggunaan dan tanpa penggunaan rubrik pada AES dan ASAG	27
8	Hasil uji signifikansi efek tetap pada model AES dan ASAG	28
9	Hasil perbandingan antarkombinasi data AES menggunakan CLD	29
10	Hasil perbandingan antarkombinasi data ASAG menggunakan CLD	30

## DAFTAR GAMBAR

1	Arsitektur <i>self-attention</i>	3
2	<i>Framework</i> penilaian otomatis berbasis LLM	6
3	Sebaran skor yang diberikan dosen pada (a) AES dan (b) ASAG	17
4	Perbandingan sebaran skor yang diberikan dosen dan LLM pada (a) AES dan (b) ASAG	18
5	Visualisasi pasangan kata yang paling sering muncul dalam teks alasan pada (a) AES dan (b) ASAG	20
6	Perbandingan kinerja LLM untuk berbagai <i>prompt</i> dalam konteks AES berdasarkan (a) nilai MSE dan (b) nilai QWK	20
7	Perbandingan kinerja LLM untuk berbagai <i>prompt</i> dalam konteks ASAG berdasarkan (a) nilai MSE dan (b) nilai QWK	21
8	Perbandingan kinerja penilaian otomatis ketika menggunakan rubrik dan tanpa menggunakan rubrik pada (a) AES dan (b) ASAG	22
9	Hasil ordinasi NMDS pada AES untuk kombinasi statistik mantel tertinggi berdasarkan (a) Gemini (b) GPT, dan (c) Claude	24
10	Hasil ordinasi NMDS pada ASAG untuk kombinasi statistik mantel tertinggi berdasarkan (a) Gemini (b) GPT, dan (c) Claude	26
11	Visualisasi interaksi efek tetap pada (a) AES dan (b) ASAG	30

## DAFTAR LAMPIRAN

1	Contoh skema <i>prompt</i> P1	36
2	Contoh skema <i>prompt</i> P5	36
3	Rubrik penilaian yang digunakan pada konteks AES	36
4	Rubrik penilaian yang digunakan pada konteks ASAG	36
5	Diagnostik model pada konteks AES	37
6	Diagnostik model pada konteks ASAG	37