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

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.

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.

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

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

Judul Skripsi : Analisis Performa *Large Language Model* dalam Penilaian Otomatis Menggunakan Model Campuran Linear
Nama : Dalilah Husna
NIM : G1401221003

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

DAFTAR GAMBAR

DAFTAR LAMPIRAN