

EVALUASI KINERJA *LARGE LANGUAGE MODEL* PADA *PERSONALIZED LEARNING SYSTEM* BERDASARKAN KERANGKA KOGNITIF

MUHAMMAD QUWWAMUL HAQ DJUNAID



**PROGRAM SARJANA ILMU KOMPUTER
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

PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Evaluasi Kinerja *Large Language Model* Pada *Personalized Learning System* Berdasarkan Kerangka Kognitif” 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.

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

Bogor, Juli 2026

Muhammad Quwwamul Haq Djunaid
G6401221055



ABSTRAK

MUHAMMAD QUWWAMUL HAQ DJUNAID. Evaluasi Kinerja *Large Language Model* pada *Personalized Learning System* Berdasarkan Kerangka Kognitif. Dibimbing oleh YENI HERDIYENI dan MEDRIA KUSUMA DEWI HARDHIENATA.

Perkembangan *Large Language Model* (LLM) mendorong pemanfaatannya dalam *Personalized Learning System* untuk mendukung pembelajaran yang lebih adaptif. Namun, kesesuaian respons LLM terhadap karakteristik kognitif pengguna masih belum banyak dievaluasi. Penelitian ini bertujuan mengevaluasi kinerja LLM dalam menghasilkan respons yang sesuai dengan karakteristik kognitif pengguna pada sistem LogiCT berbasis *Computational Thinking*. Penelitian menggunakan pendekatan evaluatif kuantitatif dengan metode *LLM-as-a-Judge* dan dua *Human Expert Judges*. Evaluasi dilakukan terhadap 48 respons yang mewakili delapan tipe kognitif dan enam level pedagogis. Respons dihasilkan menggunakan Gemma 3 27B dan dievaluasi oleh GPT-5.2 melalui OpenRouter API menggunakan skala 0–2. Hasil evaluasi menunjukkan bahwa kesesuaian *LLM-as-a-Judge* terhadap *Human Expert Judge* 1 dan *Human Expert Judge* 2 menghasilkan *accuracy* masing-masing sebesar 85,42% dan 68,75%, *Weighted F1-Score* sebesar 0,82 dan 0,73, *Weighted Kappa* sebesar 0,17 dan 0,14, serta Gwet's AC2 sebesar 0,96 dan 0,85. Hasil tersebut menunjukkan bahwa LogiCT mampu menghasilkan respons yang sesuai dengan karakteristik kognitif pengguna.

Kata kunci: *Computational Thinking*, Kerangka kognitif, *Large Language Model*, *LLM-as-a-Judge*, *Personalized Learning System*.

ABSTRACT

MUHAMMAD QUWWAMUL HAQ DJUNAID. Performance Evaluation of Large Language Models in Personalized Learning Systems Based on a Cognitive Framework. Supervised by YENI HERDIYENI and MEDRIA KUSUMA DEWI HARDHIENATA.

The rapid advancement of Large Language Models (LLMs) has driven their adoption in Personalized Learning Systems to support more adaptive learning. However, the alignment of LLM-generated responses with users' cognitive characteristics has received limited attention. This study evaluates the performance of LLM-generated responses in the LogiCT system for Computational Thinking using a quantitative evaluative approach with an *LLM-as-a-Judge* and two *Human Expert Judges*. A total of 48 responses representing eight cognitive types across six pedagogical levels were generated using Gemma 3 27B and evaluated by GPT-5.2 via the OpenRouter API on a 0–2 scale. The agreement between the *LLM-as-a-Judge* and *Human Expert Judges* achieved *accuracy* values of 85.42% and 68.75%, *Weighted F1-Scores* of 0.82 and 0.73, *Weighted Kappa* values of 0.17 and 0.14, and Gwet's AC2 values of 0.96 and 0.85, respectively. These findings indicate that LogiCT generally produces responses aligned with users' cognitive characteristics.

Keywords: *Cognitive framework*, *Computational Thinking*, *Large Language Model*, *LLM-as-a-Judge*, *Personalized Learning System*.



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 2026¹
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.

EVALUASI KINERJA *LARGE LANGUAGE MODEL* PADA *PERSONALIZED LEARNING SYSTEM* BERDASARKAN KERANGKA KOGNITIF

MUHAMMAD QUWWAMUL HAQ DJUNAID

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Sarjana Ilmu Komputer

**PROGRAM SARJANA ILMU KOMPUTER
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:

1 Rina Trisminingsih, S.Kom., M.T., Ph.D.



@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 : Evaluasi Kinerja *Large Language Model* pada *Personalized Learning System* Berdasarkan Kerangka Kognitif

Nama : Muhammad Quwwamul Haq Djunaid

NIM : G6401221055

Disetujui oleh

Pembimbing 1:

Prof. Dr. Yeni Herdiyeni, S.Si., M.Kom.



Pembimbing 2:

Dr. Medria Kusuma Dewi Hardhienata,
S.Kom., Ph.D.



Diketahui oleh

Ketua Program Sarjana Ilmu Komputer:

Dr. Sony Hartono Wijaya, S.Kom., M.Kom.

NIP 19810809 200812 1 002



Tanggal Ujian:
14 Juli 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. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Agustus 2025 sampai bulan Juli 2026 ialah Kecerdasan Buatan pada Pendidikan, dengan judul “Evaluasi Kinerja *Large Language Model* pada *Personalized Learning System* Berdasarkan Kerangka Kognitif”. Penulis menyampaikan terima kasih kepada:

1. Prof. Dr. Yeni Herdiyeni, S.Si., M.Kom. dan Dr. Medria Kusuma Dewi Hardhienata, S.Kom., Ph.D. selaku dosen pembimbing yang telah memberikan arahan dan bimbingan selama proses penyusunan karya ilmiah ini.
2. Bapak/Ibu dosen, moderator seminar, dan penguji yang telah memberikan beragam kontribusi demi penyempurnaan karya ilmiah ini.
3. Ayah, ibu, dan seluruh keluarga tercinta atas doa dan dukungan selama menempuh pendidikan hingga penyelesaian karya ilmiah ini.
4. Shyfa, Felix, Firoos, Agal, Maysa, Rio, Fadhil, Aji, Mocha, Yudha, dan Zhaki selaku rekan *Tim Personalized Learning System for Computational Thinking (PLS for CT)* yang telah berjuang bersama menyelesaikan penelitian ini.
5. Annas, Deu, Feng, Rana, dan Oce yang telah menemani dari awal MPKMB hingga berjuang bersama di tahun ke-4 untuk menyelesaikan tugas akhir S1.
6. Dicky, Nabila, Namira, Talenta, Rere, Ardayan, Muti, dan Windy yang telah menemani semenjak PKU dan membantu penulis merapihkan karya ilmiah ini.
7. Teman-teman Ilmu Komputer ST33, teman KKN, serta teman intern yang telah hadir mendukung dan turut membantu penulis menyelesaikan karya ilmiah ini.
8. Shyfa, selaku rekan penelitian pada PLS for CT yang telah menjadi teman diskusi serta eksekusi dalam segala kondisi demi kelancaran penelitian ini.
9. Seluruh pihak yang tidak dapat disebutkan satu per satu yang telah memberikan bantuan dan dukungan hingga penyelesaian karya ilmiah ini.

Semoga karya ilmiah ini dapat memberikan manfaat bagi pihak yang membutuhkan dan berkontribusi bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2026

Muhammad Quwwamul Haq Djunaid

DAFTAR ISI

DAFTAR TABEL	x
DAFTAR GAMBAR	x
DAFTAR LAMPIRAN	xi
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
1.4 Manfaat	2
1.5 Ruang Lingkup	3
II TINJAUAN PUSTAKA	5
2.1 Kerangka Kognitif Dalam Pendidikan	5
2.2 <i>Large Language Model</i>	6
2.3 <i>Large Language Model</i> dalam Kerangka Kognitif	7
2.4 <i>Personalized Learning System</i>	7
2.5 <i>Computational Thinking</i>	8
III METODE	9
3.1 Data Penelitian	9
3.2 Tahapan Penelitian	11
3.3 Peralatan Penelitian	18
IV HASIL DAN PEMBAHASAN	21
4.1 Generasi Respons LLM LogiCT	21
4.2 Hasil Evaluasi oleh <i>LLM-as-a-Judge</i>	22
4.3 Hasil Evaluasi oleh <i>Human Expert Judges</i>	24
4.4 Komparasi Hasil Evaluasi antara <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i>	26
V SIMPULAN DAN SARAN	47
5.1 Simpulan	47
5.2 Saran	47
DAFTAR PUSTAKA	49
LAMPIRAN	52
RIWAYAT HIDUP	66



DAFTAR TABEL

1	Karakteristik pada tiap kategori soal	9
2	Definisi tiap tipe profil	10
3	Struktur <i>dataset</i>	10
4	Kode tiap tipe kognitif	12
5	Skala penilaian	15
6	Contoh hasil generasi respons	21
7	Hasil evaluasi kesesuaian penilaian antara <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> secara keseluruhan	42
8	Hasil evaluasi kesesuaian penilaian antara <i>LLM-as-a-Judge</i> dan <i>Human Expert Judge 1</i> pada tiap tipe kognitif	42
9	Hasil evaluasi kesesuaian penilaian antara <i>LLM-as-a-Judge</i> dan <i>Human Expert Judge 2</i> pada tiap tipe kognitif	43
10	Hasil evaluasi kesesuaian penilaian antara <i>Human Expert Judge 1</i> dan <i>Human Expert Judge 2</i> pada tiap tipe kognitif	44

DAFTAR GAMBAR

1	Kerangka kognitif hasil adaptasi Dunn & Dunn (1975)	5
2	Tahapan penelitian	11
3	Distribusi skor hasil evaluasi oleh <i>LLM-as-a-Judge</i>	22
4	Rataan skor hasil evaluasi oleh <i>LLM-as-a-Judge</i> pada tiap tipe kognitif	23
5	Distribusi skor hasil evaluasi oleh <i>Human Expert Judges</i>	24
6	Rataan skor hasil evaluasi oleh <i>Human Expert Judges</i> pada tiap tipe kognitif	25
7	Perbandingan distribusi skor hasil evaluasi	27
8	Perbandingan rataan skor hasil evaluasi pada tiap tipe kognitif	28
9	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil TGI	29
10	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil TGR	31
11	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil TAI	32
12	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil TAR	33
13	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil PGI	34
14	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil PGR	35
15	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil PAI	36
16	Skor yang diberikan oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i> pada tiap respons dengan profil PAR	38
17	Distribusi skor dibawah 2 tiap evaluator berdasarkan tingkatan soal	40



DAFTAR LAMPIRAN

1	<i>Chain-of-Thought</i> yang disusun	53
2	<i>System prompt</i> yang disusun untuk melakukan generasi respons	54
3	<i>System prompt</i> yang disusun untuk melakukan evaluasi respons	55
4	Kode perhitungan analisis kuantitatif	58
5	Kode perhitungan analisis kesesuaian	61
6	Hasil evaluasi oleh <i>LLM-as-a-Judge</i> dan <i>Human Expert Judges</i>	64



@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