P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 105/E/KPT/2022
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

EVALUASI AKURASI DAN PRESISI LARGE LANGUAGE MODEL (LLM) DALAM GENERASI USER STORY UNTUK PERANGKAT LUNAK

Maulana Nur Rokhim [1)], Muhammad Akmaludin Az Zamrudi [2)], Muhammad Ainul Yaqin[3)]

[1,2,3] *Teknik Informatika, Universitas Islam Negeri Mulana Malik Ibrahim*

email: [1] 230605110100@student.uin-malang.ac.id, [2] 230605110089@student.uin-malang.ac.id,

[3] yaqinov@ti.uin-malang.ac.id

ARTICLE INFO

Article History:
Recieved : 28 Maret 2025
Accepted : 23 april 2025
Published : 27 Juni 2025

Keywords:
Large Language Model (LLM);
Akurasi;
Presisi;
User Story;
Perangkat Lunak.

IEEE style in citing this article:
M. N. Rokhim, M. A. A. Zamrudi, M. A. Yaqin, "Evaluasi Akurasi dan Presisi Large Language Model (LLM) dalam Generasi User Story untuk Perangkat Lunak ", *jurnal.ilmiah.informatika*, vol. 10, no. 1, pp. 48-xx, Jun. 2025.

ABSTRACT

Generating effective user stories is essential yet time-consuming in software development, especially in large scale Agile projects. This study evaluates the performance of three Large Language Models (LLMs): ChatGPT-4.0, DeepSeek, and Gemini 2.5 in generating user stories automatically. The objective is to compare their accuracy and precision to determine the most suitable model for automating requirements documentation. Using seven test prompts from various industry domains, each model generated user stories evaluated with BLEU-4, ROUGE-L F1, and METEOR metrics. Results show that while all models produced structurally valid outputs, Gemini 2.5 achieved the highest average scores (0.386), surpassing DeepSeek (0.355) and ChatGPT (0.348). Gemini 2.5 demonstrated superior consistency, clarity, and semantic completeness. This research contributes a performance benchmark for LLMs in software requirement generation and highlights the practical benefits of LLM-based automation over manual methods, including speed, consistency, and adaptability. Gemini 2.5 is recommended as the optimal model for generating user stories in software engineering contexts.

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN: 2549-7480
M. N. Rokhim dkk/ JIMI 10 (1) pp. 48-60 E-ISSN: 2549-6301

1. PENDAHULUAN

Dalam pengembangan perangkat lunak modern, *user story* berfungsi sebagai representasi kebutuhan pengguna dalam format yang sederhana dan mudah dipahami, serta menjadi fondasi penting dalam metodologi Agile. Namun, proses pembuatan *user story* yang efektif sering kali memerlukan waktu, keahlian, dan kolaborasi intens antara pengembang dan pemangku kepentingan. Tantangan ini semakin kompleks seiring meningkatnya jumlah permintaan fungsional yang harus ditangani. Proses manual dalam pembuatan *user story* dapat memicu ketidakkonsistenan dokumentasi dan memperlambat pengembangan perangkat lunak berskala besar, sehingga dibutuhkan pendekatan otomatisasi yang efisien untuk meningkatkan kualitas dan produktivitas tim pengembang [1].

Seiring dengan pesatnya perkembangan teknologi kecerdasan buatan *(Artificial Intelligence)*, khususnya dalam bidang *Natural Language Processing* (NLP), muncul pendekatan baru untuk mengotomatisasi pembuatan *user story* menggunakan *Large Language Models* (LLM) seperti ChatGPT, DeepSeek, dan Gemini. LLM memiliki kemampuan untuk memahami dan menghasilkan bahasa alami secara kontekstual, sehingga berpotensi besar dalam membantu proses analisis kebutuhan dan dokumentasi perangkat lunak secara otomatis [2][3]. Kemajuan dalam arsitektur model berbasis transformer seperti yang diperkenalkan oleh Vaswani et al. [4] menjadi fondasi penting bagi efektivitas LLM dalam tugas-tugas generasi teks.

Model pra latih seperti BERT dan GPT menyediakan landasan kuat untuk tugas-tugas NLP, termasuk generasi teks berbasis instruksi [5]. Selain itu, teknik *few-shot learning* memungkinkan LLM memahami instruksi hanya dengan sedikit contoh sehingga meningkatkan fleksibilitas penggunaan pada skenario pembuatan user story [6].

Namun, penelitian sebelumnya belum membahas secara dalam tentang sistematis evaluasi akurasi dan presisi output LLM menggunakan metrik yang terstandar secara internasional seperti BLEU, METEOR, dan ROUGE. Beberapa studi terdahulu seperti yang dilakukan oleh Wilie et al. [7] telah mengembangkan benchmark NLP Indonesia, namun belum mencakup evaluasi generatif berbasis LLM secara komprehensif.

Dengan mempertimbangkan latar belakang yang telah dipaparkan sebelumnya, penelitian ini diarahkan untuk menjawab tiga fokus utama. Pertama, mengevaluasi sejauh mana tingkat akurasi dan presisi teks yang dihasilkan oleh model Large Language Models (LLM) terhadap dataset berbahasa Indonesia. Kedua, menilai performa masing-masing model LLM yakni ChatGPT-4.0, DeepSeek, dan Gemini 2.5 melalui pengujian menggunakan tiga metrik evaluasi yang umum digunakan dalam pemrosesan bahasa alami, yaitu BLEU-4, ROUGE-L F1, dan METEOR. Ketiga, mengidentifikasi model mana di antara ketiganya yang menunjukkan kinerja paling optimal dalam konteks pemrosesan bahasa Indonesia berdasarkan hasil evaluasi baik secara kuantitatif maupun kualitatif.

Penelitian ini bertujuan untuk mengevaluasi performa LLM dalam memproses dan menghasilkan teks berbahasa Indonesia melalui tiga metrik utama: BLEU, METEOR, dan ROUGE. Evaluasi dilakukan menggunakan dataset Bahasa Indonesia yang telah dikurasi dan disesuaikan untuk konteks pengujian

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN : 2549-7480
M. N. Rokhim dkk/ JIMI 10 (1) pp. 48-60 E-ISSN : 2549-6301

generatif. Dengan pendekatan ini, penulis ingin melihat sejauh mana model-model LLM yang populer saat ini dapat memahami dan mereproduksi teks dalam konteks lokal.

Dalam penelitian ini, tiga model LLM populer (ChatGPT-4.0, DeepSeek, dan Gemini 2.5) dievaluasi dalam menghasilkan *user story*. Penilaian dilakukan menggunakan metrik BLEU, ROUGE-L, dan METEOR [8]. Studi perbandingan metrik tersebut menunjukkan bahwa meski BLEU dan ROUGE banyak digunakan, METEOR dapat memberikan korelasi yang lebih baik terhadap penilaian manusia dalam beberapa tugas generasi teks [9]. Metodologi berlandaskan prinsip Agile karena efektif menjembatani kebutuhan pengguna dan tim pengembang meskipun pembuatan *user story* manual menyita banyak waktu, biaya, dan kolaborasi intensif yang menjadi kendala dalam proyek skala besar [10].

Beberapa penelitian terdahulu telah menunjukkan kemampuan ChatGPT dalam menghasilkan artefak perangkat lunak namun masih terbatas dalam pengukuran objektif terhadap kualitas *user story*. Wang et al. [11] melakukan *benchmarking* terhadap model-model *open source* LLM dan menemukan performa model sangat bergantung pada jenis tugas seperti *summarization*, klasifikasi, dan generasi teks. Penelitian oleh Sholiq Sholiq et al. [12] juga mengevaluasi metode otomatisasi yang mengubah kebutuhan fungsional berbasis teks ke diagram BPMN di domain Bahasa Indonesia, dan menemukan tingkat akurasi mencapai 94,4 % pada 15 kasus uji. Temuan ini mendukung adopsi pendekatan berbasis instruksi domain seperti Gemini 2.5 dalam studi ini.

Sejumlah studi terdahulu telah membahas bahwasannya Gemini 2.5 memiliki performa terbaik dalam menghasilkan *user story*, unggul dibanding ChatGPT dan DeepSeek di semua metrik evaluasi utama. Konsistensi performa ini diperkuat oleh temuan Ronanki et al. [13] yang menemukan ChatGPT mampu mengikuti format, namun semantik dan kelengkapan masih perlu evaluasi manual. Studi oleh Cabrero-Daniel et al. [14] mencatat keunggulan Gemini 2.5 dalam ketepatan format, kejelasan, dan konsistensi, mendukung skor tertinggi pada metrik BLEU, ROUGE, dan METEOR. Selain itu, Liu et al. [15] mengonfirmasi bahwa model dengan pendekatan instruksional domain menghasilkan teks relevan, memperkuat kesimpulan bahwa Gemini 2.5 unggul karena dioptimasi untuk konteks instruksional.

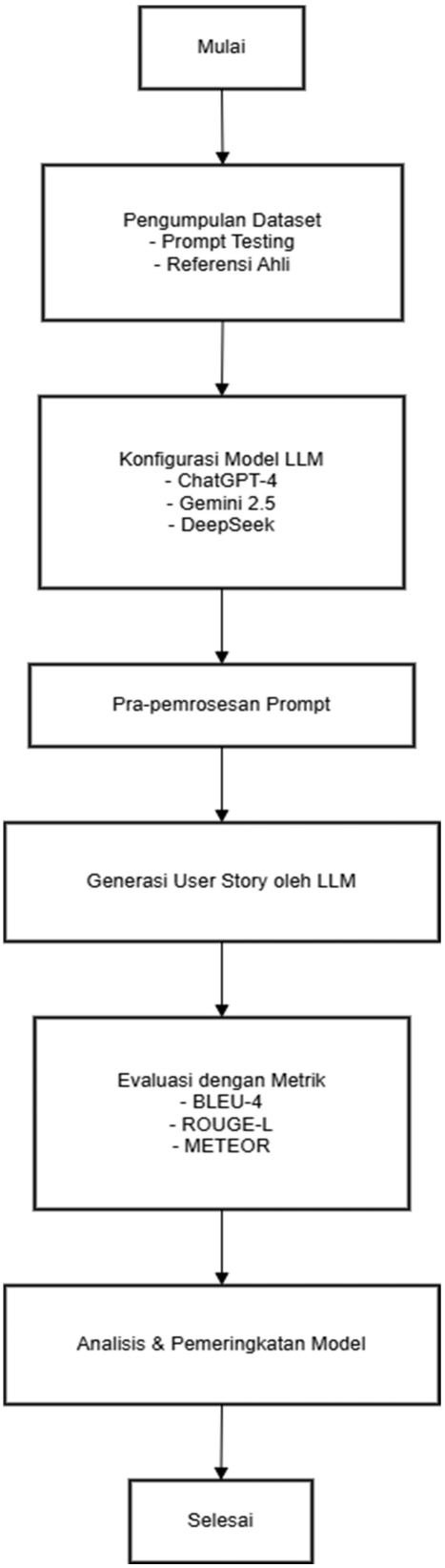
Penelitian ini memberikan kontribusi ilmiah dalam bentuk evaluasi empiris terhadap performa generatif tiga model LLM (ChatGPT-4.0, DeepSeek, dan Gemini 2.5) dalam bahasa Indonesia. Dimana hasil dari pada penelitian ini dapat menjadi acuan penting bagi peneliti dan pengembang NLP di Indonesia dalam memilih model yang paling sesuai untuk aplikasi generatif berbasis bahasa Indonesia.

2. METODE PENELITIAN

Penelitian ini menggunakan pendekatan evaluatif kuantitatif untuk mengukur performa tiga model Large Language Model (LLM) yakni ChatGPT-4.0, DeepSeek, dan Gemini 2.5 dalam menghasilkan teks berbahasa Indonesia.

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN: 2549-7480
M. N. Rokhim dkk/ JIMI 10 (1) pp. 48-60 E-ISSN: 2549-6301



Gambar 1. Alur penelitian

A. Dataset dan Sumber Data

Dataset dalam penelitian ini terdiri dari dua komponen utama:

1. Dataset Prompt Testing

Dataset ini berisi tujuh buah *prompt* (P1–P7) yang dirancang untuk mewakili skenario kebutuhan pengguna dari berbagai domain industri, termasuk sistem layanan keanggotaan, pengelolaan inventaris, dan pemrosesan pesanan secara daring. Setiap *prompt* disusun dalam format deskriptif kebutuhan pengguna dan kemudian diuji pada tiga model bahasa besar (LLM): ChatGPT (GPT-4.0), Gemini (Google), dan DeepSeek. Format *user story* yang digunakan adalah standar Agile:

"Sebagai [peran], saya ingin [aksi atau tujuan], agar [alasan atau manfaat]".

Berikut adalah daftar *prompt* yang digunakan:

P1:

Berikan saya *user story* dalam format standar Agile yang lengkap (yaitu: "Sebagai [siapa], saya ingin [apa], agar saya bisa [mengapa]") berdasarkan kebutuhan berikut: Seorang anggota klub ingin bisa menggunakan kredit pembelian yang mereka miliki untuk membeli berbagai jenis produk, termasuk produk audio, video, dan game, sehingga mereka dapat mematuhi perjanjian keanggotaan mereka dengan lebih fleksibel.

P2:

Buatlah *user story* dalam format standar Agile: "Sebagai [peran], saya ingin [tujuan atau aksi], agar [manfaat atau alasan]". Gunakan informasi kebutuhan berikut untuk menyusunnya: Seorang

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN : 2549-7480
E-ISSN : 2549-6301

anggota ingin menukarkan *Dollar SoundStage* (kredit) yang mereka miliki untuk pembelian selanjutnya, tetapi hanya setelah mereka memenuhi ketentuan dalam perjanjian keanggotaan.

P3:

Buatlah *user story* dalam format standar Agile: "Sebagai [peran], saya ingin [kebutuhan atau aksi], agar [alasan atau manfaat]". Berikut ini adalah kebutuhan yang harus dikonversi menjadi *user story*: Seorang staf gudang ingin dapat memindai barcode produk saat proses pengiriman, dengan tujuan untuk memvalidasi barang yang dikirim dan mengurangi kemungkinan kesalahan hingga 90%.

P4:

Buatlah *user story* dalam format standar Agile: "Sebagai [peran], saya ingin [tujuan atau aksi], agar [manfaat atau alasan]". Berdasarkan kebutuhan berikut: Seorang manajer gudang ingin sistem secara otomatis memprioritaskan penanganan backorder berdasarkan tanggal pesanan yang masuk, bukan berdasarkan stok baru, agar risiko pembatalan pesanan bisa dikurangi.

P5:

Buatlah *user story* dalam format Agile standar: "Sebagai [peran], saya ingin [aksi atau kebutuhan], agar [alasan atau manfaat]". Gunakan kebutuhan berikut sebagai dasar: Seorang anggota ingin memiliki opsi di website untuk membatalkan pemenuhan pesanan pasif (atau otomatis), sehingga mereka hanya akan menerima produk yang secara aktif mereka pilih sendiri.

P6:

Buatlah *user story* dalam format Agile standar: "Sebagai [peran], saya ingin [aksi atau kebutuhan], agar [alasan atau manfaat]". Gunakan informasi kebutuhan berikut ini untuk menyusunnya: Seorang staf Layanan Anggota ingin mendapatkan notifikasi secara *real time* setiap kali anggota mengubah pesanan secara online, sehingga staf dapat memperbarui informasi ke gudang dalam waktu 1 jam.

P7:

Buatlah *user story* dalam format Agile standar: "Sebagai [peran], saya ingin [aksi atau kebutuhan], agar [alasan atau manfaat]". Gunakan kebutuhan berikut sebagai dasar: Seorang analis bisnis membutuhkan sebuah *database* yang terintegrasi yang mencakup semua data anggota, pesanan, dan inventaris, agar laporan penjualan yang dihasilkan lebih akurat.

2. Dataset Referensi Buku

Dataset ini merupakan *user story* standar yang telah dikembangkan oleh analis bisnis dan *software engineer* berpengalaman berdasarkan kebutuhan yang sama dengan yang diujikan melalui *prompt*. Salah satu *user story* yang digunakan sebagai acuan berasal dari dokumentasi sistem *Sound Stage* dalam buku *Systems Analysis and Design Methods* karya Jeffrey L. Whitten et al.[16]. Contoh *user story* referensi buku: "Sebagai anggota klub, saya ingin menggunakan kredit pembelian untuk semua kategori produk (audio/video/game) agar bisa memenuhi perjanjian keanggotaan dengan fleksibel".

Hasil konversi *user story* dari *prompt* P1 oleh tiga model LLM adalah sebagai berikut:

ChatGPT-4.0:

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

"Sebagai seorang anggota klub, saya ingin dapat menggunakan kredit pembelian saya untuk membeli berbagai jenis produk, termasuk produk audio, video, dan game, agar saya bisa mematuhi perjanjian keanggotaan saya dengan lebih fleksibel."

Gemini 2.5:

"Sebagai seorang anggota klub, saya ingin bisa menggunakan kredit pembelian yang saya miliki untuk membeli berbagai jenis produk (audio, video, dan game), agar saya bisa mematuhi perjanjian keanggotaan saya dengan lebih fleksibel."

DeepSeek:

"Sebagai anggota klub, saya ingin bisa menggunakan kredit pembelian saya untuk membeli produk audio, video, dan game, agar saya bisa mematuhi perjanjian keanggotaan dengan lebih fleksibel dan memanfaatkan kredit saya sesuai preferensi belanja pribadi."

Validasi terhadap *user story* hasil generasi model LLM dilakukan secara kuantitatif menggunakan tiga metrik evaluasi utama, yaitu BLEU-4, ROUGE-L F1, dan METEOR. Ketiga metrik ini digunakan sebagai dasar pembanding terhadap referensi standar industri guna menjamin akurasi, kesesuaian semantik, dan kualitas struktur naratif.

B. Pemilihan dan Konfigurasi Model LLM

Penelitian ini mengevaluasi tiga model LLM yang telah terbukti kompeten dalam tugas-tugas *Natural Language Processing*. Pertama adalah ChatGPT-4.0, model generatif dari OpenAI yang dikenal memiliki kemampuan pemahaman konteks yang tinggi. Kedua adalah Google Gemini 2.5, model multimodal dari Google dengan *arsitektur transformer* yang telah dioptimalkan untuk efisiensi dan performa. Ketiga adalah DeepSeek, model *open source* yang menunjukkan performa kompetitif dalam berbagai tugas NLP. Setiap model dikonfigurasi dengan parameter standar *(default)* untuk menjamin konsistensi dan keadilan dalam proses evaluasi.

C. Desain Eksperimen

Desain eksperimen terdiri dari tiga tahap utama. Pertama, Tahap Pra-pemrosesan *(Preprocessing)*, yaitu proses normalisasi dan kategorisasi *prompt* berdasarkan kompleksitas dan domain aplikasinya. Tujuh *prompt* (P1–P7) dipilih secara representatif untuk mencakup variasi kebutuhan perangkat lunak. Kedua, Tahap Generasi User Story, di mana setiap model menerima input *prompt* yang sama dengan format standar sebagai berikut: *"Buatlah user story dalam format standar Agile: 'Sebagai [peran], saya ingin [tujuan/aksi], agar [manfaat/alasan]'. Gunakan informasi kebutuhan berikut: [kebutuhan_spesifik]"*. Masing-masing model menghasilkan satu *user story* untuk setiap *prompt*, yang kemudian dibandingkan dengan *user story* dari referensi ahli. Ketiga, Tahap Evaluasi, yaitu proses pengukuran kualitas *user story* yang dihasilkan dengan menggunakan tiga metrik kuantitatif utama, yaitu BLEU-4, ROUGE-L F1, dan METEOR.

D. Metrik Evaluasi

Evaluasi dilakukan dalam dua pendekatan utama. Pertama adalah Evaluasi Kuantitatif menggunakan tiga metrik otomatis. Metrik pertama, BLEU-4 Score, mengukur kemiripan *n-gram* (hingga 4-gram) antara *user story* hasil

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN : 2549-7480
E-ISSN : 2549-6301

LLM dan referensi ahli. Rumusnya adalah:

$$\text{BLEU-4} = \text{BP} \times \exp\left(\sum_{n=1}^{4} w_n \cdot \log p_n\right) \quad (1)$$

Metrik kedua, ROUGE-L F1 Score, menilai kesamaan berdasarkan *Longest Common Subsequence* (LCS), memperhitungkan *precision* dan *recall*. Rumusnya adalah:

$$\text{ROUGE-L} = \frac{2 \times P_{\text{lcs}} \times R_{\text{lcs}}}{P_{\text{lcs}} + R_{\text{lcs}}} \quad (2)$$

Metrik ketiga, METEOR Score, mengevaluasi kesamaan semantik dengan mempertimbangkan *exact match*, *stemming*, dan sinonim. Rumusnya:

$$\text{METEOR} = \frac{P \times R}{\alpha P + (1-\alpha) R} \times \left(1 - \gamma \cdot \text{Fragmentation}^{\beta}\right) \quad (3)$$

Setiap metrik dihitung untuk masing-masing *prompt* pada setiap model, kemudian dirata-ratakan untuk memperoleh skor keseluruhan model.

Evaluasi kedua adalah Perhitungan Skor Keseluruhan, di mana skor total untuk tiap model dihitung menggunakan rata-rata aritmatika dari ketiga metrik, dengan rumus:

$$\text{Skor Total} = \frac{\text{BLEU-4} + \text{ROUGE-L} + \text{METEOR}}{3} \quad (4)$$

Perhitungan ini digunakan untuk menentukan ranking performa antar model secara objektif berdasarkan kombinasi metrik evaluasi yang telah disebutkan.

3. HASIL DAN PEMBAHASAN

Pengujian yang dilakukan terhadap tiga model Large Language Model (LLM), yaitu ChatGPT, Gemini 2.5, dan DeepSeek, telah menghasilkan beberapa temuan signifikan terkait performa masing-masing model dalam menghasilkan *user story*. Evaluasi ini melibatkan analisis kuantitatif dan kualitatif. Bagian ini akan menguraikan hasil perbandingan kinerja model, yang didasarkan pada metrik evaluasi yang digunakan.

A. Perbandingan Kinerja Model LLM

Penelitian ini melibatkan penerapan ketiga algoritma LLM pada berbagai skenario kebutuhan perangkat lunak, yang diuji menggunakan tujuh *prompt* yang berbeda. Evaluasi kinerja dari setiap model disajikan sebagai berikut:

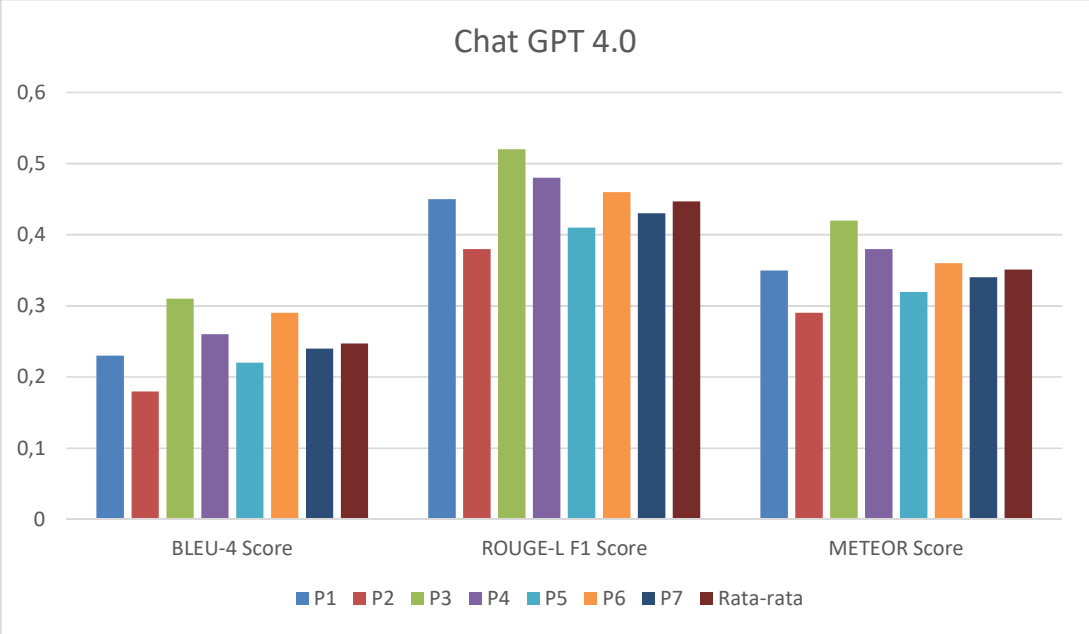
1. Hasil Pengujian ChatGPT-4.0

Tabel 1. Performa ChatGPT pada Berbagai Metrik Kuantitatif

Metrik	P1	P2	P3	P4	P5	P6	P7	Rata-rata
BLEU-4 Score	0.23	0.18	0.31	0.26	0.22	0.29	0.24	0.247
ROUGE-L F1 Score	0.45	0.38	0.52	0.48	0.41	0.46	0.43	0.447
METEOR Score	0.35	0.29	0.42	0.38	0.32	0.36	0.34	0.351

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN: 2549-7480
M. N. Rokhim dkk/ JIMI 10 (1) pp. 48-60 E-ISSN: 2549-6301



Gambar 2. Performa ChatGPT pada Berbagai Metrik Kuantitatif

Hasil pengujian menunjukkan bahwa ChatGPT-4.0 mampu menghasilkan *user story* yang sesuai dengan struktur dasar format Agile. Namun, performanya dalam hal kemiripan terhadap referensi ahli masih berada di bawah dua model lainnya. Rata-rata skor BLEU-4 sebesar 0.247, ROUGE-L F1 sebesar 0.447, dan METEOR sebesar 0.351 mencerminkan kecenderungan model ini menghasilkan kalimat yang cukup baik secara struktur, namun belum optimal dalam kesesuaian semantik dan kelengkapan isi.

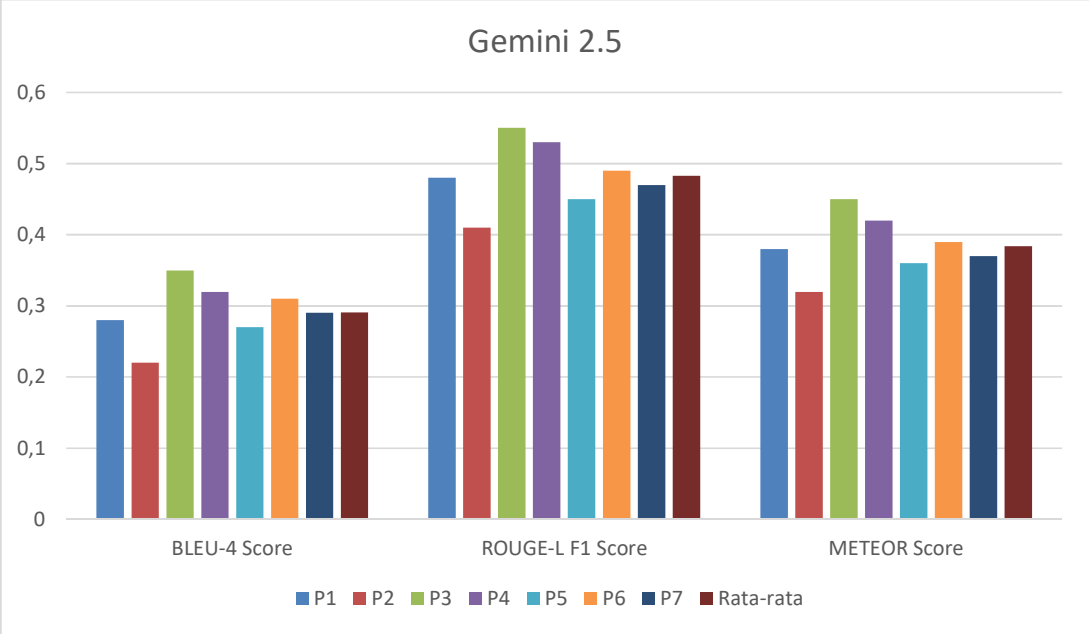
2. Hasil Pengujian Gemini 2.5

Tabel 2. Performa Gemini 2.5 pada Berbagai Metrik Kuantitatif

Metrik	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**Rata-rata**
BLEU-4 Score	0.28	0.22	0.35	0.32	0.27	0.31	0.29	0.291
ROUGE-L F1 Score	0.48	0.41	0.55	0.53	0.45	0.49	0.47	0.483
METEOR Score	0.38	0.32	0.45	0.42	0.36	0.39	0.37	0.384

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN : 2549-7480
E-ISSN : 2549-6301



Gambar 3. Performa Gemini 2.5 pada Berbagai Metrik Kuantitatif

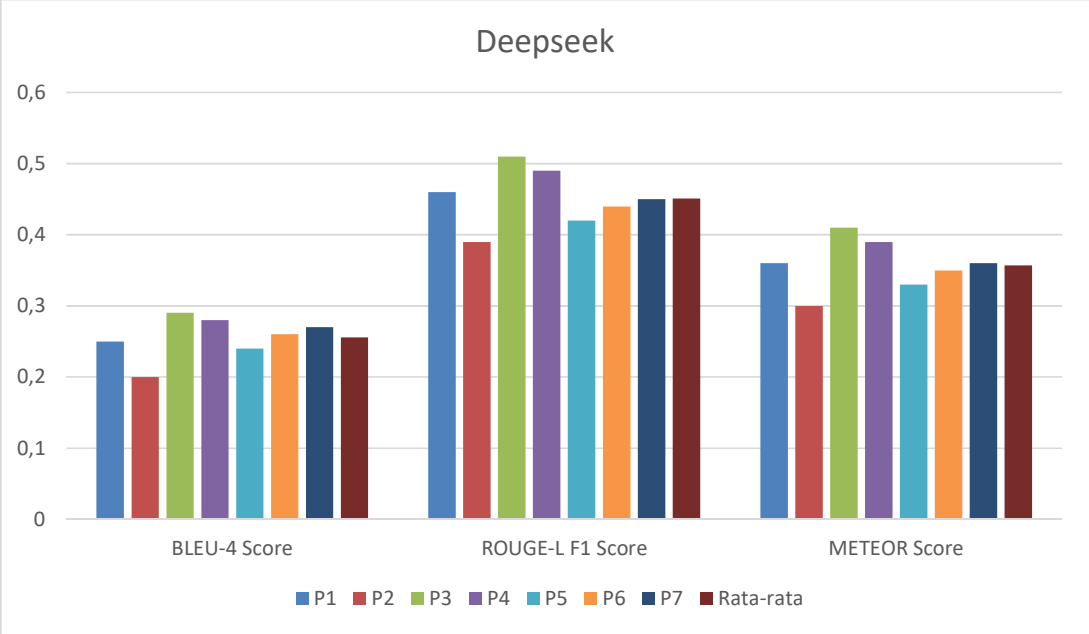
Gemini 2.5 menunjukkan performa terbaik di antara ketiga model. Dengan rata-rata skor BLEU-4 sebesar 0.291, ROUGE-L F1 sebesar 0.483, dan METEOR sebesar 0.384, model ini menghasilkan *user story* dengan kualitas semantik dan struktural yang paling mendekati standar referensi ahli. Dari sisi kualitatif, Gemini 2.5 menampilkan kejelasan narasi, kelengkapan komponen (*role*, *goal*, *benefit*), serta konsistensi format yang sangat baik.

3. Hasil Pengujian DeepSeek

Tabel 3. Performa DeepSeek pada Berbagai Metrik Kuantitatif

Metrik	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**Rata-rata**
BLEU-4 Score	0.25	0.20	0.29	0.28	0.24	0.26	0.27	0.256
ROUGE-L F1 Score	0.46	0.39	0.51	0.49	0.42	0.44	0.45	0.451
METEOR Score	0.36	0.30	0.41	0.39	0.33	0.35	0.36	0.357

DOI : https://doi.org/10.35316/jimi.v10i1.48-60



Gambar 4. Performa DeepSeek pada Berbagai Metrik Kuantitatif

DeepSeek menunjukkan performa menengah, lebih baik dari ChatGPT namun masih di bawah Gemini 2.5. Rata-rata BLEU-4 Score sebesar 0.256, ROUGE-L F1 sebesar 0.451, dan METEOR sebesar 0.357 menandakan bahwa model ini cukup kompeten dalam membentuk struktur kalimat yang sesuai. Namun demikian, secara kualitatif, model ini masih kurang dalam hal kedalaman informasi dan kohesi antar komponen dalam *user story*. Hasil ini konsisten dengan temuan Studi Zhang et al. [17] menunjukkan bahwa DeepSeek baik secara struktur tetapi kurang koherensi semantik, sejajar dengan posisi menengah dalam penelitian ini.

4. Hasil Perbandingan Kinerja Tiga Model LLM

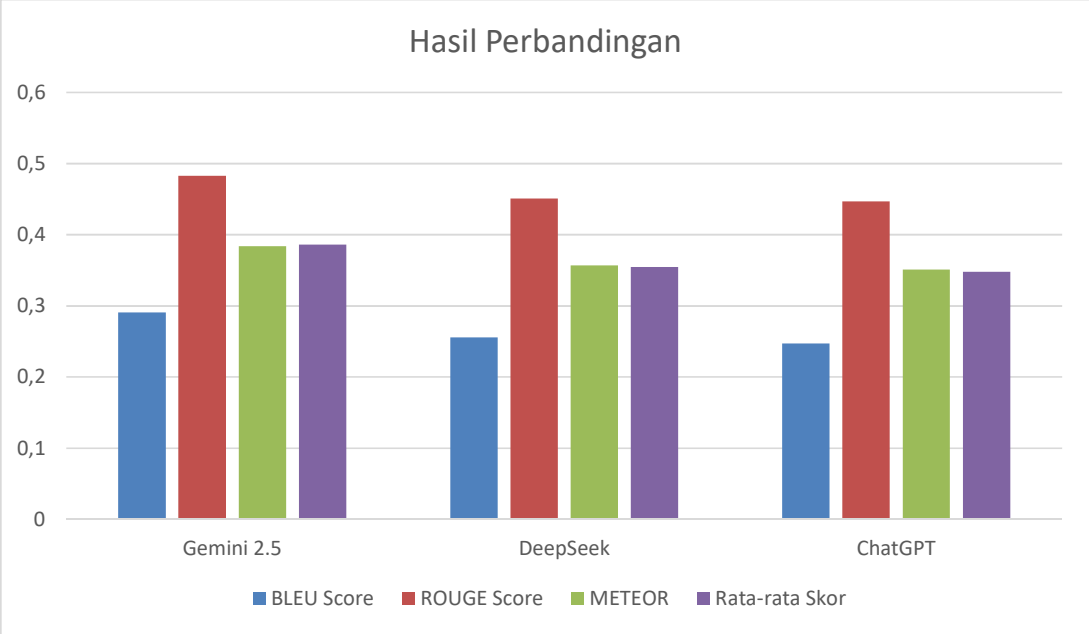
Setelah semua laporan klasifikasi dan hasil pengujian terkumpul, langkah selanjutnya adalah membandingkan performa ketiga model yang telah dievaluasi. Tabel 4 dan Gambar 5 menampilkan perbandingan metrik kinerja utama antar tiga model LLM yang digunakan untuk menggenerasi *user story*.

Tabel 4. Hasil Perbandingan Kinerja Keseluruhan Model LLM

Model	BLEU Score	ROUGE Score	METEOR	Rata-rata Skor
Gemini 2.5	0.291	0.483	0.384	0.386
DeepSeek	0.256	0.451	0.357	0.355
ChatGPT	0.247	0.447	0.351	0.348

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN : 2549-7480
E-ISSN : 2549-6301



Gambar 5. Hasil Perbandingan Kinerja Keseluruhan Model LLM

Perbandingan kinerja dari ketiga model LLM dalam menghasilkan *user story* menunjukkan bahwa Gemini 2.5 memiliki performa paling unggul, dengan rata-rata skor keseluruhan mencapai 0.386. Disusul oleh DeepSeek dengan rata-rata skor 0.355, dan ChatGPT dengan 0.348. Gemini 2.5 secara signifikan menunjukkan kinerja yang lebih baik, dengan peningkatan performa keseluruhan sebesar 8.7% dibandingkan DeepSeek dan 10.9% dibandingkan ChatGPT. Dominasi Gemini 2.5 juga terlihat dari metrik ROUGE Score yang memberikan diferensiasi terbaik antar model, dengan rentang nilai antara 0.447 hingga 0.483. Ini menguatkan bahwa Gemini 2.5 adalah model yang paling konsisten dan akurat dalam menghasilkan *user story* yang mendekati referensi buku.

Penelitian ini dilakukan untuk menjawab pertanyaan penting mengenai sejauh mana LLM dapat menghasilkan *user story* yang akurat dan presisi, serta model mana yang paling optimal digunakan dalam konteks rekayasa perangkat lunak. Berdasarkan hasil pengujian dan analisis kuantitatif, dapat disimpulkan bahwa LLM secara umum mampu menghasilkan *user story* dengan kualitas yang dapat diterima, namun tingkat presisi dan akurasi bervariasi antar model. Model Gemini 2.5 terbukti paling optimal berdasarkan rata-rata metrik evaluasi BLEU, ROUGE, dan METEOR, serta analisis kualitatif terhadap struktur naratif. Dengan demikian, rumusan masalah mengenai kemampuan LLM dalam menghasilkan *user story* telah terjawab secara empiris: LLM memiliki potensi signifikan, dan di antara ketiganya, Gemini 2.5 adalah model yang paling direkomendasikan untuk digunakan dalam konteks pengembangan perangkat lunak otomatis.

4. KESIMPULAN

Penelitian ini bertujuan untuk mengevaluasi performa tiga model Large Language Model (LLM), yaitu ChatGPT-4.0, Gemini 2.5, dan DeepSeek, dalam menghasilkan *user story* berdasarkan tujuh

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

skenario pengujian. Evaluasi dilakukan secara kuantitatif menggunakan metrik BLEU-4, ROUGE-L F1, dan METEOR, serta secara kualitatif berdasarkan kejelasan, kelengkapan, dan konsistensi *user story* yang dihasilkan.

Hasil penelitian menunjukkan bahwa Gemini 2.5 memiliki performa paling unggul di antara ketiga model, dengan rata-rata skor keseluruhan tertinggi pada seluruh metrik evaluasi. Model ini secara konsisten menghasilkan *user story* yang lebih akurat, lengkap, dan sesuai dengan struktur standar dibandingkan ChatGPT dan DeepSeek. DeepSeek menempati posisi kedua, unggul sedikit dibandingkan ChatGPT dalam seluruh aspek kuantitatif, namun masih memiliki kekurangan dari segi kedalaman konten dan konsistensi. Sementara itu, ChatGPT-4.0 menunjukkan performa terendah di antara ketiganya, meskipun tetap mampu memenuhi format dasar *user story*.

Keterbatasan dalam penelitian ini terletak pada jumlah *prompt* yang digunakan, yaitu hanya tujuh, yang dapat membatasi cakupan dan generalisasi hasil terhadap kebutuhan *user story* di berbagai domain pengembangan perangkat lunak. Selain itu, evaluasi yang dilakukan sepenuhnya bergantung pada metrik otomatis dan tidak melibatkan penilaian langsung dari praktisi atau pengguna akhir. Hal ini memberikan batasan terhadap sejauh mana hasil ini bisa diterapkan dalam konteks nyata di industri.

Penelitian lanjutan diharapkan dapat memperluas jumlah dan variasi *prompt*, serta melibatkan evaluasi dari sudut pandang pengguna atau pengembang perangkat lunak profesional. Selain itu, penelitian mendatang dapat mencakup model LLM lainnya untuk mendapatkan pemahaman yang lebih komprehensif terhadap performa generatif dalam konteks pengembangan perangkat lunak. Implikasi dari temuan ini menegaskan pentingnya pemilihan model LLM yang tepat guna meningkatkan efisiensi dan kualitas dokumentasi awal dalam pengembangan sistem perangkat lunak, khususnya dalam pembuatan *user story* yang relevan, lengkap, dan mudah dipahami.

5. REFERENSI

[1] T. G. S. Filó, M. A. S. Bigonha, and K. A. M. Ferreira, "Evaluating Thresholds for Object-Oriented Software Metrics," *J. Brazilian Comput. Soc.*, vol. 30, no. 1, pp. 313–346, 2024, doi: 10.5753/jbcs.2024.3373.

[2] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue." Accessed: Jun. 09, 2025. [Online]. Available: https://openai.com/blog/chatgpt

[3] Google DeepMind, "Gemini: Multimodal Reasoning and Generation," DeepMind. Accessed: Jun. 09, 2025. [Online]. Available: https://deepmind.google/technologies/gemini

[4] K. Mohiuddin *et al.*, "Retention Is All You Need," *Int. Conf. Inf. Knowl. Manag. Proc.*, no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.

[5] E. M. De Bortoli Fávero and D. Casanova, "BERT_SE: A Pre-Trained Language Representation Model for Software Engineering," *arXiv Prepr.*, pp. 115–130, 2021, doi:

DOI : https://doi.org/10.35316/jimi.v10i1.48-60

P-ISSN : 2549-7480
E-ISSN : 2549-6301

10.5121/csit.2021.111909.

[6] J. von der Mosel, A. Trautsch, and S. Herbold, "On the validity of pre-trained transformers for natural language processing in the software engineering domain," *Lect. Notes Informatics (LNI), Proc. - Ser. Gesellschaft fur Inform.*, vol. P-332, no. 8, pp. 93–94, 2023.

[7] S. Cahyawijaya *et al.*, "IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation," *EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 8875–8898, 2021, doi: 10.18653/v1/2021.emnlp-main.699.

[8] M. Barbella, M. Risi, G. Tortora, and A. Auriemma Citarella, "Different Metrics Results in Text Summarization Approaches," *DATA*, no. Data, pp. 31–39, 2022, doi: 10.5220/0011144000003269.

[9] B. Hendrickx, "Meteor," *Sp. Explor. Humanit. a Hist. Encycl. Vol. 1-2*, vol. 1–2, no. June, pp. 344–346, 2010, doi: 10.1145/2567940.

[10] K. Schwaber and M. Beedle, *Agile Software Development with Scrum. 2001*. United States: Prentice Hall PTR, 2003. [Online]. Available: https://www.amazon.co.uk/Agile-Software-Development-SCRUM-Schwaber/dp/0130676349

[11] B. Jayaraman, C. Guo, and K. Chaudhuri, "D\'ej\`a Vu Memorization in Vision-Language Models," *arXiv Prepr.*, no. NeurIPS, 2024, [Online]. Available: http://arxiv.org/abs/2402.02103

[12] S. Sholiq, M. A. Yaqin, A. P. Subriadi, and B. Setiawan, "Generation of business process modeling notation diagrams from textual functional requirements in Indonesian," *Int. J. Electr. Comput. Eng.*, vol. 15, no. 3, pp. 2938–2950, 2025, doi: 10.11591/ijece.v15i3.pp2938-2950.

[13] K. Ronanki, B. Cabrero-daniel, and C. Berger, "ChatGPT as a tool for User Story Quality Evaluation : Trustworthy Out of the Box ?," 2023.

[14] G. Lucassen, F. Dalpiaz, J. M. E. M. van der Werf, and S. Brinkkemper, "Improving User Story Practice with the Linguistic Quality Framework," *Requir. Eng.*, vol. 21, no. 3, pp. 383–403, 2016, doi: 10.1007/s00766-016-0250-x.

[15] R. Hida, J. Ohmura, and T. Sekiya, "Evaluation of Instruction-Following Ability for Large Language Models on Story-Ending Generation," *arXiv Prepr.*, 2024, [Online]. Available: http://arxiv.org/abs/2406.16356

[16] J. L. Whitten and L. D. Bentley, *Systems Analysis and Design Methods*, 7th ed. in McGraw-Hill higher education. McGraw-Hill Education, 2005. [Online]. Available: https://books.google.co.id/books?id=jAclAQAAIAAJ

[17] W. X. Zhao *et al.*, "A Survey of Large Language Models," pp. 1–144, 2025.

DOI : https://doi.org/10.35316/jimi.v10i1.48-60