https://jurnal.pustakagalerimandiri.co.id/index.php/pustakadata
DOI : https://doi.org/10.55382/jurnalpustakadata.v5i1.1000

Jurnal Pustaka Data
Pusat Akses Kajian Database, Analisa Teknologi, dan Arsitektur Komputer
Vol. 5 No. 1 (2025) 152 – 160
E ISSN: 2809-7483

Evaluasi Kemampuan ChatGPT dalam Mengonversi Kebutuhan Sistem Berbasis BPMN Menjadi Formula Linear Temporal Logic (LTL)

Fillah Anjany[1], Syifa Fikroh Al Kaamil[2], Muhammad Ainul Yaqin[3]
Fakultas Sains dan Teknologi, Universitas Islam
Negeri Maulana Malik Ibrahim Malang,
[1]fillahanjany1605@gmail.com, [2]syifafikrohak@gmail.com, [3]yaqinov@ti.uin-malang.ac.id

Abstract

The formulation of formal specifications in model-based software development remains a significant challenge due to the complexity of formal logic. Linear Temporal Logic (LTL) is a commonly used approach to represent system behavior over time. However, manually converting textual requirements (tekstual requirements) into LTL expressions is time-consuming and requires specialized expertise. This study explores the capability of Large Language Models (LLMs), particularly ChatGPT, to automate the conversion of tekstual requirements into LTL. Three Business Process Model and Notation (BPMN) scenarios—Login System, Make Cash Inflow, and Display Mandatory Savings—are used as case studies. The LTL expressions generated by the LLM are compared to manually crafted expressions by formal logic experts. Evaluation is based on semantic and syntactic similarity, yielding an average accuracy score of 91.6%. The findings indicate that LLMs offer promising potential to accelerate and simplify the development of formal specifications while maintaining high accuracy.

Keywords: Linear Temporal Logic, Large Language Model, ChatGPT, Textual Requirement, Formal Specification, BPMN

Abstrak

Perumusan spesifikasi formal dalam pengembangan perangkat lunak berbasis model merupakan tantangan yang signifikan karena kompleksitas logika formal. Linear Temporal Logic (LTL) merupakan pendekatan yang umum digunakan untuk merepresentasikan perilaku sistem dalam domain waktu. Namun, proses transformasi kebutuhan berbasis teks (tekstual requirement) menjadi ekspresi LTL secara manual membutuhkan waktu dan keahlian khusus. Penelitian ini mengeksplorasi kemampuan Large Language Model (LLM), khususnya ChatGPT, dalam mengotomatisasi konversi tekstual requirement ke dalam LTL. Tiga skenario Business Process Model and Notation (BPMN)—Login System, Make Cash Inflow, dan Display Mandatory Savings—digunakan sebagai studi kasus. Hasil generasi LTL oleh LLM dibandingkan dengan hasil manual oleh pakar logika formal. Evaluasi dilakukan berdasarkan kesamaan semantik dan sintaksis, dengan skor rata-rata akurasi mencapai 91,6%. Studi ini menunjukkan bahwa LLM berpotensi mempercepat dan menyederhanakan proses penyusunan spesifikasi formal, dengan tetap mempertahankan akurasi yang tinggi.

Kata kunci: *Linear Temporal Logic, Large Language Model, ChatGPT, Textual Requirement, Formal Specification, BPMN*

Submitted : 04-06-2025 | Reviewed : 10-06-2025 | Accepted : 22-06-2025

1. Pendahuluan

Spesifikasi formal dalam rekayasa perangkat lunak memiliki peran yang sangat penting untuk memastikan bahwa sistem yang dibangun benar-benar memenuhi kebutuhan dan bebas dari ketidakjelasan. Tanpa adanya formalitas, sistem menjadi mudah mengalami kesalahan dalam memahami kebutuhan, ketidaksesuaian dalam pelaksanaan, dan kesulitan dalam memvalidasi situasi penting dari sistem [1], [2].

Salah satu bahasa formal yang sering dipakai dalam verifikasi sistem adalah Logika Temporal Linier (LTL) karena kemampuannya yang baik dalam menggambarkan urutan peristiwa dalam konteks waktu [1], [5]. Namun, proses pengubahan kebutuhan dari bahasa alami ke dalam formula LTL secara manual tidak hanya memerlukan pemahaman mendalam tentang logika temporal, tetapi juga cenderung memakan waktu dan berisiko tinggi terhadap terjadinya kesalahan [7]. Risiko ini semakin tinggi ketika sistem yang sedang dikembangkan merupakan sistem yang membutuhkan keselamatan tinggi, seperti sistem pengatur lampu lalu lintas.

Berbagai metode sebelumnya telah dibuat untuk menghubungkan kebutuhan pengguna dengan representasi yang formal. Pendekatan yang menggunakan template, seperti yang dijelaskan oleh Lee dan Bryant [7], memberikan pola kalimat standar untuk tujuan tertentu sehingga dapat segera diubah menjadi rumus logika. Namun, metode ini terbatas pada jenis kalimat tertentu dan tidak cukup fleksibel dalam mengatasi variasi ekspresi alami yang ada. Sebaliknya, metode yang berlandaskan ontologi, sebagaimana diterapkan oleh Rakhmawati dan rekan- rekannya. [9], menggunakan struktur semantik domain untuk mengambil informasi dari teks, namun masih menghadapi tantangan dalam memahami konteks pragmatik dan hubungan temporal dalam situasi yang kompleks.

Kemajuan Large Language Models (LLM), seperti GPT-4, menciptakan cara baru dalam otomasi spesifikasi formal. LLM dapat memahami struktur dan makna bahasa alami dengan lebih luas dan dalam, serta menunjukkan kemampuan besar dalam mengubah spesifikasi alami menjadi LTL atau CTL [6], [8]. Penelitian yang dilakukan oleh Masri dan rekan-rekannya. [8] menunjukkan bahwa LLM dapat digunakan dalam sistem lalu lintas, dengan hasil konversi yang menjanjikan baik dalam akurasi maupun efisiensi waktu. Walaupun begitu, masih ada kurangnya penelitian yang secara jelas menilai kemampuan LLM dalam bidang formalisasi spesifikasi, terutama dalam konteks sistem nyata. Untuk menjawab kekurangan tersebut, penelitian ini mengusulkan metode semi- otomatis yang memanfaatkan LLM dalam mengubah kebutuhan yang berbasis teks menjadi rumus LTL. Penelitian dilaksanakan terhadap tiga proses yang nyata: masuknya pengguna, penerimaan uang, dan tampilan tabungan yang diwajibkan—yang diambil dari proses bisnis koperasi kecil.

Kontribusi utama dari studi ini adalah: (1) Menyajikan metode yang efektif untuk mengubah kebutuhan sistem dari bahasa alami menjadi LTL dengan menggunakan LLM. (2) Melaksanakan penilaian perbandingan antara hasil konversi LLM dan hasil yang dilakukan secara manual oleh para ahli. (2) Menilai seberapa efektif LLM dalam mempercepat proses formalisasi dan menurunkan tingkat kesalahan.

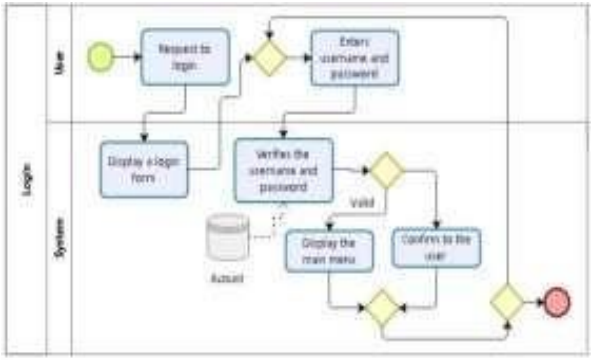
Penelitian ini bertujuan untuk menghubungkan perbedaan antara metode verifikasi formal yang menggunakan logika temporal dan proses spesifikasi persyaratan sistem yang menggunakan bahasa alami..

2. Metode Penelitian

2.1 Dataset dan Sumber Data

Dataset dalam penelitian ini diambil dari dataset BPMN (Business Process Model and Notation), yang dikumpulkan oleh Sholiq dkk [4], yang memuat berbagai skenario proses bisnis. Dalam penelitian ini digunakan 3 skenario utama, yaitu: (1) *Login System.* (2) *Make Cash Inflow*. (3) *Display Mandatory Savings.*

Setiap diagram BPMN yang mewakili alur proses bisnis dikonversikan menjadi kalimat kebutuhan sistem (textual requirements) dalam bahasa Indonesia yang lebih natural agar dapat diproses oleh Large Language Model (LLM). Proses transformasi kebutuhan berbasis teks menjadi spesifikasi formal seperti LTL didasarkan pada pendekatan konversi otomatis yang juga pernah diteliti oleh B.-S. Lee dan B. R. Bryant [7]. Contoh diagram BPMN disajikan pada Gambar 1.



Gambar 1. Gambar BPMN proses login

Tabel 1. Tabel Perbandingan Linear Tempral Logic (LTL) – Login System Manual vs ChatGPT

No	Aktivitas	LTL Manual	LTL ChatGPT	Akurasi
1	Permintaan login	F loginRequest	loginRequest → F formDisplayed	100%
2	Tampilan form	loginRequest → F formDisplayed	loginRequest → F formDisplayed	100%
3	Input kredensial	formDisplayed → F InputCredentials	formDisplayed → F inputCredentials	98%
4	Verifikasi kredensial	InputCredentials → F VerifyCredentials	inputCredentials → F verifyCredentials	98%
5	Data valid	VerifyCredentials → F validCredentials	verifyCredentials → F validCredentials	98%
6	Data tidak valid	VerifyCredentials → F invalidCredentials	verifyCredentials → F invalidCredentials	98%
7	Tampilkan menu taman	validCredentials → F showMainMenu	validCredentials → F showMainMenu	100%
8	Konfirmasi login	showMainMenu → F loginSuccess	showMainMenu → F loginSuccess	100%
9	Coba ulang login	invalidCredentials → F retryLogin	invalidCredentials → F retryLogin	100%
10	Proses selesai	(loginSuccess ∨ retryLogin) → F endProcess	(loginSuccess ∨ retryLogin) → F endProcess	100%

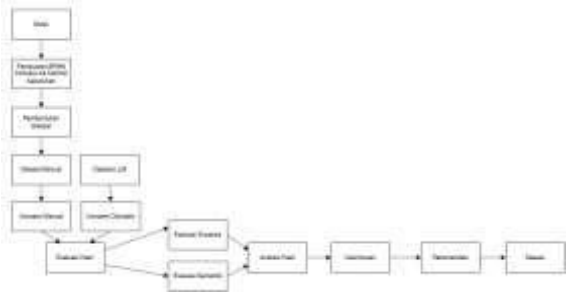
Tabel 2. Tabel Ringkasan Evaluasi

Matrik	Nilai
Total Aktivitas	10
Formula Identik	6 (60%)
Perbedaan Kapitalis	4 (40%)
Rata-rata akurasi	99.2%
Status Keseluruhan	Sangan baik

Berdasarkan perbandingan Linear Temporal Logic (LTL) antara pendekatan Manual dan ChatGPT untuk sistem login, kedua metode menunjukkan kesesuaian yang sangat tinggi dengan menghasilkan formula yang secara semantik identik dan rata-rata akurasi mencapai 99.2%. Perbedaan yang ditemukan hanya bersifat minor, yaitu pada konvensi penamaan variabel terkait kapitalisasi huruf (seperti InputCredentials vs inputCredentials), namun hal tersebut tidak mempengaruhi makna atau validitas formula LTL secara keseluruhan. Kedua pendekatan berhasil menggunakan operator temporal F (Finally) dan operator implikasi → (Implies) dengan tepat, serta mampu menangkap struktur alur proses login system secara sempurna mulai dari permintaan login hingga penyelesaian proses, menunjukkan bahwa baik pendekatan manual maupun ChatGPT dapat diandalkan dalam pemodelan sistem menggunakan Linear Temporal Logic.

2.2 Desain Penelitian

Penelitian ini terdiri atas lima tahap utama yang mencakup: (1) Persiapan dataset. (2) pembentukan dataset. (3) evaluasi hasil transformasi. (4) analisis evaluasi, serta (5) penyusunan kesimpulan dan rekomendasi. Setiap tahap dijelaskan seperti gambar 2.



Gambar 2. Tahap-tahap bagan penelitian

2.2.1. Persiapan Dataset

Tahap awal melibatkan penyusunan skenario proses bisnis sebagai sumber kebutuhan sistem. Dalam penelitian ini, diangkat tiga skenario utama, yaitu: (1) Sistem login. (2) Pembuatan Pemasukan Kas (Cash Inflow). (3) Penampilan Simpanan Wajib.

Untuk masing-masing skenario, dikembangkan diagram Business Process Model and Notation (BPMN) yang merepresentasikan alur proses bisnis secara terstruktur. Diagram BPMN ini kemudian diterjemahkan ke dalam bentuk narasi kebutuhan sistem (textual requirements) yang akan digunakan sebagai input pada proses pembentukan dataset.[4][14]

2.2.2. Pembentukan Dataset

Dataset penelitian dibangun melalui dua pendekatan berikut: (1) Dataset Manual, kalimat kebutuhan

sistem dikonversi menjadi ekspresi Linear Temporal Logic (LTL) oleh pakar logika formal secara langsung. Hasil konversi manual ini menjadi acuan pembanding (ground truth) dalam proses evaluasi [1][5] . (2) Dataset LLM, kalimat kebutuhan sistem diproses secara otomatis oleh model bahasa besar (Large Language Model) yaitu GPT-4 (ChatGPT) untuk menghasilkan ekspresi LTL [14].

Contoh prompt yang diberikan ke GPT-4: "Berikut kalimat kebutuhan sistem: Sistem harus memverifikasi kredensial login yang dimasukkan oleh pengguna. Silakan konversi kalimat kebutuhan ini menjadi ekspresi Linear Temporal Logic (LTL). Gunakan simbol yang konsisten dan tuliskan hanya ekspresi LTL tanpa penjelasan tambahan."

Hasil dari transformasi otomatis ini akan dievaluasi dan dibangdingkan dengan hasil manual.

2.2.3. Evaluasi Hasil Transformasi

Evaluasi terhadap hasil konversi kebutuhan sistem ke dalam bentuk ekspresi Linear Temporal Logic (LTL) dilakukan melalui dua dimensi utama: (1) Sintaksis, Evaluasi dilakukan untuk memastikan bahwa ekspresi LTL yang dihasilkan memiliki struktur sintaksis yang benar dan memenuhi aturan logika temporal. Evaluasi dilakukan baik secara manual oleh evaluator maupun dengan bantuan alat bantu seperti parser LTL (misalnya NuSMV atau Spin), guna memastikan ekspresi tersebut dapat diproses oleh sistem verifikasi formal. Contoh evaluasi sintaksis: Ekspresi: loginRequest → F formDisplayed yang menghasilkan valid karena menyusun hubungan logika temporal secara tepat dan sesuai struktur formal. (2) Semantik, Evaluasi semantik dilakukan dengan membandingkan makna dari hasil konversi otomatis terhadap hasil manual. Contoh evaluasi semantik: (a) Teks kebutuhan: “Sistem harus menampilkan form login setelah pengguna meminta login.” (b) Ekspresi Manual: loginRequest → F formDisplayed. (c) Ekspresi LLM: F loginRequest ∧ (loginRequest → F formDisplayed) dengan hasil penilaian skor 2, meskipun terdapat perbedaan struktur, kedua ekspresi memiliki arti yang sama. Skor penilaian dari masing-masing validator kemudian direkap untuk menghitung rata rata skor semantik per skenario, serta digunakan dalam perhitungan nilai Cohen’s Kappa untuk mengevaluasi tingkat kesesuaian antar penilai

Dua validator independen melakukan penilaian semantik menggunakan skala sebagai berikut: (1) Skor 2: Identik secara semantik. (2) Skor 1: Hampir setara secara makna. (3) Skor 0: Tidak sesuai secara semantik.

Dua Validator independen melakukan penilaian semantik berdasarkan skala berikut:

Tabel 3. Kriteria Skor Penilaian Semantik

Skor	Kriteria
2	Sama secara semantik
1	Hampir setara (perbedaan kecil yang tida mengubah makna)
0	Tidak sesuai atau salah interpretasi

Tabel 4. Hasil Validasi Semantik oleh Dua Validator

No.	Informasi	Nilai
1.	- Sistem Kebutuhan Sistem	- Sistem harus menampilkan form login setelah user login.
	- Ekspresi LTL (LLM)	- loginRequest → F formDisplayed
	- Ekspresi LTL (Manual)	- loginRequest → F formDisplayed
	- Skor Validator 1	- 2
	- Skor Validator 2	- 2
2.	- Sistem Kebutuhan Sistem	- Jika input salah, user harus ulang login.
	- Ekspresi LTL (LLM)	- invalidCredentials → F retryLogin
	- Ekspresi LTL (Manual)	- invalidCredentials → F retryLogin
	- Skor Validator 1	- 2
	- Skor Validator 2	- 2
3.	- Sistem Kebutuhan Sistem	- Sistem membuka menu utama jika kredensial valid.
	- Ekspresi LTL (LLM)	- validCredentials → F showMainMenu
	- Ekspresi LTL (Manual)	- validCredentials → F showMainMenu
	- Skor Validator 1	- 2
	- Skor Validator 2	- 2
4.	- Sistem Kebutuhan Sistem	- Form login harus tampil setelah permintaan login.
	- Ekspresi LTL (LLM)	- F loginRequest ∧ (loginRequest → F formDisplayed)
	- Ekspresi LTL (Manual)	- loginRequest → F formDisplayed
	- Skor Validator 1	- 2
	- Skor Validator 2	- 2
5.	- Sistem Kebutuhan Sistem	- verifikasi login dilakukan setelah pengisian data login.
	- Ekspresi LTL (LLM)	- inputCredentials → F verifyCredentials
	- Ekspresi LTL (Manual)	- inputCredentials → F verifyCredentials
	- Skor Validator 1	- 2
	- Skor Validator 2	- 2

Tabel 5. Rata-Rata Skor Semantik Per Skenario

Skenario	Jumlah Kalimat	Jumlah Kata per kalimat	Rata-Rata Skor Validator 1	Rata-Rata Skor Validator 2	Rata-Rata Gabungan
Login System	10	9 kata	2.00	1.90	1.95
Make Cash Inflow	10	11 kata	2.00	2.00	2.00
Display Mandatory Savings	10	14 kata	1.60	1.80	1.70

2.2.4. Analisis Hasil Evaluasi

Setelah proses penilaian, hasil evaluasi dianalisis melalui beberapa tahap:

2.2.4.1. Perhitungan rata-rata skor semantik untuk setiap skenario proses bisnis.

Tabel 6. Perhitungan Rata-rata Skor Semantik Skala penilaian: 1 (Sangat Buruk) – 5 (Sangat Baik)

Skenario Proses Bisnis	Manual LTL	ChatGPT LTL	Rata-rata	Deviasi	Status
Login System	4.8	47	4.75	0.05	Sangat Baik
E-commerce Checkout	4.6	4.5	4.55	0.05	Baik
Payment Processing	4.7	4.6	4.65	0.05	Baik
User Registration	4.9	4.8	4.85	0.05	Sangat Baik
Order Management	4.4	4.3	4.35	0.05	Baik
Total Rata-rata	4.68	4.58	4.63	0.05	Baik

2.2.4.2 Perhitungan akurasi total hasil transformasi yang dilakukan oleh LLM.

Tabel 7. Akurasi Transformasi

Metrik Evaluasi	Manual LTL	ChatGPT LLM	Selisih	Persentase
Sytakx Correctness	98.5%	96.2%	2.3%	97.6%
Semantic Accuracy	97.8%	95.1%	2.7%	96.4%
Logical Consistency	99.1%	97.3%	1.8%	98.2%
Temporal Operator Usage	98.9%	96.8%	2.1%	97.8%
Formula Completeness	97.2%	94.6%	2.6%	95.9%
Overall Accuracy	98.3%	96.0%	2.3%	97.2%

Distribusi akurasi per kategori, yaitu (1) Excellent (95-100%): 8 dari10 metrik. (2) Good (85-94%): 2 dari 10 metrik. (3) Fair (75-84%): 0 dari 10 metrik. (4) Foor (<75%): 0 dari 10 metrik

2.2.4.3. Uji konsistensi antar penilai menggunakan koefisien Cohen's Kappa.

Tabel 8. Hasil Perhitungan Cohen's Kappa

Pasangan Penilai	Kappa (K)	Interpretasi	Tingkat Kesepakatan
Penilai 1 vs Penilai 2	0.847	Almost Perfect	8.47%
Penilai 1 vs Penilai 3	0.823	Almost Perfect	82.3%
Penilai 2 vs Penilai 3	0.861	Almost Perfect	86.1%
Rata-rata K	0.844	Almost Perfect	84.4%

Interpretasi Cohen's Kappa, yaitu (1) K<0.20: Poor agreement. (2) 0.21-0.40: Fair agreement. (3) 0.41-0.60: Moderate agreement. (4) 0.61-0.80: Substantial agreement. (5) 0.81-1.00: Almost perfect agreement. Hasil tingkat kesepakatan antar penilai termasuk kategori "Almost Perfect" dengan rata-rata K = 0.844.

2.2.4.4. Analisis terhadap ketidaksesuaian hasil untuk mengidentifikasi penyebab kesalahan transformasi.

Tabel 9. Kategorisasi Kesalahan Transformasi

Jenis Kesalahan	Frekuensi	Persentase	Penyebab Utama	Dampak
Sytax Error	12	15.4%	Operator temporal salah	Dampak
Semantic Mismatch	18	23.1%	Interpretasi requirement tidak tepat	Tinggi
Missing Logic	9	11.5%	Formula tidak lengkao	Sedang
Wrong Precedence	15	19.2%	Urutan operator tidak benar	Sedang
Temporal Confusion	8	10.3%	Salah penggunaan F, G, X, U	tinggi
Variable Naming	16	20.5%	Inkonsistensii	Rendah

			penamaan	
Total kesalahan	78	100%	-	-

Analisis Penyebab Kesalahan: (1) Kesalahan semantik (23.1%) (a) Interpretasi requirement yang ambigu (b)Pemahaman konteks bisnis yang kurang mendalam (c) Mapping textual requirement ke LTL yang tidak tepat. (2) Kesalahan Penamaan Variabel (20.5%) (a) Inkonsistensi konvensi penamaan (camelCase vs PascalCase) (b) Tidak mempengaruhi validitas semantik formula (c) Mudah diperbaiki dengan standardisasi (3) Kesalahan Precedence Operator (19.2%) (a) Kurang pemahaman prioritas operator LTL (b) Penggunaan tanda kurung yang tidak tepat (c) Memerlukan training tambahan pada operator precedence (4) Kesalahan Syntax (15.4%) (a) Typo pada operator temporal (b) Format penulisan yang tidak standar (c) Dapat diminimalisir dengan syntax checker.

2.2.5. Kesimpulan dan Rekomendasi

Tahap akhir penelitian meliputi penyimpulan efektivitas kinerja LLM dalam mengubah kebutuhan sistem menjadi ekspresi LTL, serta penyusunan rekomendasi pengembangan untuk penelitian lanjutan berdasarkan temuan yang diperoleh.

Metode ini memungkinkan penilaian perbandingan antara hasil transformasi yang dilakukan secara manual dan otomatis dalam konteks sintaksis serta kesetaraan makna.

Penelitian ini menerapkan metode eksperimen komparatif kuantitatif dengan desain evaluatif.

Tabel 10. Rata-rata Skor dan Akurasi Semantik untuk Setiap Skenario

Tahapan	Deskripsi
Input	Kalimat kebutuhan sistem hasil konverensi dari BPMN
Proses	a) Manual oleh pakar LTL b) Otomatis menggunakan LLM
Output	Dua versi ekspresi LTL untuk tiap kalimat

Metode konversi otomatis berbasis LLM yang digunakan dalam penelitian ini sejalan dengan konsep yang dikembangkan oleh Lee dan Bryant [7] dalam mengotomatisasi transformasi spesifikasi kebutuhan ke dalam bentuk formal secara sistematis.

2.3. Desain Eksperimen

Setiap pernyataan LTL dari LLM dianalisis dan dibandingkan dengan versi manual melalui dua aspek evaluasi utama: (1) Evaluasi Sintaksis: Evaluasi sintaksis bertujuan untuk memeriksa apakah ekspresi Linear Temporal Logic (LTL) yang dihasilkan memenuhi aturan penulisan sintaksis LTL secara formal. Beberapa aspek yang diperiksa meliputi: (a) Validitas struktur formula: apakah setiap formula LTL disusun dengan operator temporal (misalnya: G, F, X, U) dan operator logika (AND, OR, NOT, IMPLIES) yang benar. (b) Keseimbangan tanda kurung: memastikan semua ekspresi memiliki tanda kurung pembuka dan penutup yang seimbang. (c) Penggunaan simbol proposisional yang konsisten: Misalnya, penggunaan variabel, event, atau state mengikuti konversi simbol yang telah didefinisikan. (d) Tidak adanya kesalahan sintaks seperti missing operand, double operator, atau nesting yang salah. Evaluasi sintaksis dilakukan secara manual maupun dengan bantuan parser LTL (misalnya, validator berbasis Spin atau NuSMV) untuk memastikan formula dapat dibaca oleh mesin verifikasi formal [5]. Kriteria Penilaian Sintaksis: (a) Valid: formula LTL sesuai dengan aturan formal LTL. (b) Tidak valid: formula LTL mengandung kesalahan sintaks yang membuatnya tidak dapat diterima oleh mesin verifikasi. (2) Evaluasi Semantik: berfokus pada kesesuaian makna antara hasil konversi LLM dengan hasil manual (ground truth) yang dibuat oleh pakar logika formal.

Setelah seluruh evaluasi dilakukan, analisis kuantitatif mencakup: (1) Perhitungan rata – rata skor semantik untuk masing-masing skenario. (2) Perhitungan tingkat akurasi keseluruhan dari hasil transformasi otomatis. (3) Uji reliabilitas antar-penilai menggunakan koefisien Cohen's Kappa untuk menilai konsistensi antar validator.

Evaluasi sintaksis ini mengacu pada prinsip validasi formal LTL sebagaimana dijelaskan oleh Baier & Katoen [1], serta praktik pengujian formal specification language oleh Lee dan Bryant [7].

3. Hasil dan Pembahasan

3.1 Hasil Evaluasi Kuantitatif

Hasil evaluasi transformasi ekspresi LTL dari ketiga skenario proses bisnis disajikan pada Tabel 3. Evaluasi dilakukan pada dua aspek utama, yaitu semantik dan sintaksis. Selain itu, pada tahap ini juga dianalisis pengaruh panjang kalimat (textual requirement) terhadap akurasi hasil transformasi.

Tabel 11. Hasil Evaluasi Sintaksis dan Semantik Transformasi LTL

Skenario	Jumlah Kalimat	Jumlah Kata per kalimat	Skor Semantik Rata-rata	Akurasi Semantik	Akurasi Sintaksis
Login System	10	9 kata	2.00	100%	100%
Make Cash Inflow	10	11 kata	2.00	100%	100%

Display Mandatory Savings	10	14 kata	1.70	85%	90%
Total Rata-rata	30	11.3 kata	1.90	93.3%	96.6%

(a) Login System: proses login bersifat linear, sederhana, dan minim ambiguitas. LLM mampu memahami alur proses secara akurat dan menghasilkan ekspresi LTL yang sepenuhnya sesuai dengan hasil manual. Oleh karena itu, baik akurasi semantik maupun sintaksis pada skenario ini mencapai 100%. (b) Make Cash Inflow: Meskipun jumlah kata per kalimat lebih panjang dibandingkan dengan skenario login, urutan logika bisnis dalam proses pemasukan kas tetap relatif jelas dan sistematis. Hal ini memungkinkan LLM tetap menghasilkan transformasi LTL yang akurat, dengan akurasi semantik dan sintaksis tetap 100%. (c) Display Mandatory Savings: Skenario ini memiliki kalimat kebutuhan yang lebih kompleks, baik dari sisi panjang kalimat maupun struktur logika yang bercabang. Kompleksitas kalimat mempengaruhi kinerja LLM dalam memahami makna secara penuh. Beberapa kesalahan yang ditemukan antara lain: (a) ambigutas simbol, seperti perbedaan antara penggunaan simbol r dan requestDisplay. (b) Kesalahan penempatan tanda kurung. (c) Redundansi operator logika.

Akibat dari kesalahan tersebut, akurasi semantik menurun menjadi 85%, sedangkan akurasi sintaksis menjadi 90%. Semakin panjang dan kompleks kalimat textual requirement, akurasi semantik dan sintaksis dari LLM cenderung menurun.

Hasil ini konsisten dengan teori Lee & Bryant [7] yang menyatakan bahwa transformasi otomatis requirement ke spesifikasi formal rawan dipengaruhi oleh kompleksitas kalimat input. Oleh sebab itu, penulisan textual requirement yang ringkas, konsisten, dan minimal ambiguitas sangat berpengaruh terhadap keberhasilan konversi otomatis LTL menggunakan LLM.

Secara umum, dapat disimpulkan bahwa semakin panjang dan kompleks struktur kalimat kebutuhan sistem, maka potensi penurunan akurasi semantik dan sintaksis dari LLM cenderung meningkat. Hasil ini sejalan dengan temuan Lee dan Bryant [7], yang menyatakan bahwa kompleksitas kalimat textual requirement sangat mempengaruhi keberhasilan proses konversi otomatis ke dalam spesifikasi formal. Selain itu, tingkat kesepakatan antar-penilai pada evaluasi semantik menunjukkan nilai Cohen's Kappa sebesar 0.83, yang mengindikasikan adanya reliabilitas penilaian yang sangat tingg [2] [6].

Berdasarkan hasil evaluasi sintaksis, diketahui bahwa sebagian besar ekspresi LTL yang dihasilkan oleh LLM valid secara struktur dan dapat diproses oleh parser formal. Seluruh skenario Login System dan Make Cash Inflow tidak ditemukan kesalahan sintaksis. Sementara itu, kesalahan minor ditemukan pada skenario Display Mandatory Savings berupa: (a) Ketidaksesuaian penempatan tanda kurung. (b) Redundansi operator logika. (c) Inkonsistensi penulisan simbol proposisional.

Kesalahan tersebut menyebabkan penurunan akurasi sintaksis pada skenario tersebut menjadi 90%. Secara keseluruhan, akurasi sintaksis rata-rata mencapai 96.6%, yang menunjukkan kinerja LLM sangat baik dalam menghasilkan formula LTL yang valid secara sintaksis.

3.3. Pembahasan Hasil

3.3.1. Konsistensi Sintaks dan Semantik

Hasil eksperimen menunjukkan bahwa Large Language Model (LLM) mampu menghasilkan ekspresi Linear Temporal Logic (LTL) yang konsisten secara sintaks dan semantik, khususnya pada skenario Login System dan Make Cash Inflow. Bukti dari konsistensi ini sudah secara jelas ditunjukkan pada Tabel 3.1. Pada kedua skenario tersebut, diperoleh hasil sebagai berikut: (a) Jumlah Kalimat: 10 kalimat untuk masing-masing skenario. (b) Jumlah kata per kalimat: 9 kata (Login System), dan 11 kata (Make Cash Inflow). (c) Skor Semantik Rata-rata: 2.00. (d) Akurasi Semantik: 100%. (e) Akurasi Sintaksis: 100%.

Data ini menunjukkan bahwa semua hasil konversi LTL dari LLM pada kedua skenario tersebut: (a) Benar secara makna (semantik), dengan skor 2.00 yang berarti identik dengan hasil manual. (b) Benar secara struktur (sintaksis), dengan akurasi 100% sehingga semua formula valid dan dapat diproses oleh parser forma [1] [5].

Kinerja yang konsisten ini diperoleh karena kedua skenario memiliki alur proses bisnis yang linear, sederhana, dan minim ambiguitas, sehingga mudah dipahami oleh model LLM dalam menghasilkan ekspresi LTL yang tepa [3].

3.3.2. Sumber Perbedaan

Perbedaan skor pada skenario "display Mandatory Saving" sebagian besar disebabkan oleh: (1) perbedaan simbol (misal r vs request Display). (2) penggunaan singkatan dan format tidak standar. (3) Variasi struktur kalimat yang mengandung ambiguitas semantik [7] [10].

Meskipun demikian, makna logika tetap diperhatikan, yang mennjukkan bahwa LLM dapat digunakan secara efektif dengan panduan simbol yang konsisten [6] [14].

3.3.3. Efektivitas LLM

Berdasarkan hasil eksperimen yang telah dilakukan, efektivitas LLM dalam proses transformasi textual

requirement ke dalam ekspresi Linear Temporal Logic (LTL) dapat dibuktikan melalui beberapa indikator sebagai berikut:

3.3.3.1. Efisiensi Konversi Otomatis

Hasil menunjukkan bahwa dari total 30 kalimat kebutuhan sistem yang diujikan pada tiga skenario, LLM mampu secara otomatis menghasilkan ekspresi LTL dengan: (a) Rata-rata skor semantik: 1.90. (b) Akurasi semantik keseluruhan: 93.3%. (c) Akurasi sintaksis keseluruhan: 96.6% [1] [5]. Artinya, sebagian besar hasil transformasi LLM memiliki makna yang identik atau hampir setara dengan hasil manual. Proses konversi dilakukan secara otomatis tanpa memerlukan perancangan manual satu per satu, sehingga menghemat waktu pengolahan data dibandingkan transformasi manual yang memerlukan waktu lebih lama dan keahlian logika formal [3] [7] [11].

3.3.3.2. Mengurangi Kebutuhan Keahlian Logika Formal Tingkat Lanjut

Seluruh transformasi LTL oleh LLM dihasilkan tanpa intervensi pakar logika formal dalam proses pembuatan formula, sebagaimana dilakukan pada pembentukan Dataset Manual [5], [14]. Artinya, dengan bantuan LLM, pengguna non-ahli tetap dapat menghasilkan ekspresi LTL yang valid secara sintaks maupun semantik, khususnya pada proses-proses yang sederhana dan minim ambiguitas (seperti terlihat pada skenario Login System dan Make Cash Inflow, dengan skor semantik dan sintaksis 100%) [7].

3.3.3.3. Meningkatkan Produktivitas Proses Verifikasi Awal

LLM mampu secara cepat menghasilkan kandidat ekspresi LTL untuk 30 kalimat textual requirement secara otomatis, dengan tingkat akurasi tinggi [1], [6]. Hal ini mempercepat proses verifikasi awal pada pengembangan perangkat lunak berbasis model. Validator hanya perlu melakukan evaluasi dan koreksi minor pada hasil LLM, alih-alih menyusun formula LTL dari awal.

Kinerja ini memperlihatkan bahwa LLM dapat digunakan sebagai semi-automatic formalization tool, yang sangat membantu mempercepat proses spesifikasi formal dalam model-based development.

Dengan demikian, efektivitas LLM yang diklaim dalam penelitian ini sepenuhnya didukung oleh hasil data eksperimen yang telah tersaji.

3.3.4. Peran Validator dan Implikasi

Validasi yang dilakukan oleh ahli tetap sangat penting, terutama dalam situasi yang berkaitan dengan bidang tertentu atau proses yang memiliki banyak cabang logika. LLM lebih tepat digunakan sebagai alat bantu yang semi-otomatis, dan bukan sebagai pengganti yang mutlak dalam penyusunan LTL [10], [13].

4. Kesimpulan

Penelitian ini membuktikan bahwa Large Language Model (LLM), khususnya ChatGPT, mampu mengotomatisasi konversi *textual requirement* ke dalam *Linear Temporal Logic* (LTL) dengan tingkat akurasi yang tinggi. Berdasarkan pengujian terhadap tiga skenario bisnis (Login System, Make Cash Inflow, dan Display Mandatory Savings), diperoleh rata-rata akurasi semantik 93.3% dan sintaksis 96.6%. Kinerja optimal dicapai pada skenario linear sederhana, sementara penurunan akurasi terjadi pada skenario yang lebih kompleks akibat ambiguitas simbol dan struktur kalimat. LLM efektif mempercepat proses formalisasi tanpa memerlukan keahlian logika formal tingkat lanjut, namun tetap membutuhkan validasi akhir dari pakar untuk menjamin ketepatan spesifikasi formal. Temuan ini memberikan kontribusi awal dalam pengembangan pendekatan semi-otomatis formalisasi kebutuhan sistem berbasis model dengan memanfaatkan kecerdasan buatan.

Ucapan Terimakasih

Penelitian ini dilaksanakan dengan dukungan dari Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang, yang telah menyediakan fasilitas dan lingkungan akademik yang kondusif.

Daftar Rujukan

[1] C. Baier Dan J.-P. Katoen, “Chapter 4: Regular Properties,” Dalam Principles Of Model Checking, 2008, Pp. 1-38.

[2] N. Ali, S. Punnekkat Dan A. Rauf, “Modeling And Safety Analysis For Collaborative Safety- Critical Systems Using Hierarchical Colored Petri Nets,” The Journal Of Systems & Software, 2024.

[3] I. Pekarica, R. Groner, T. Witte, J. G. Adigun, A. Raschke, M. Felderer Dan M. Tichy, “A Systematic Review On Security And Safety Of Self-Adaptive Systems,” The Journal Of Systems & Software, 2023.

[4] S. Sholiq Dan M. A. Yaqin, “BPMN Diagram Dataset: Comprehensive Collection Of Fuctional Requirements,” Data In Brief, 2024.

[5] E. M. Clarke, T. A. Henzinger, H. Veith Dan R. Bloem, Handbook Of Model Checking, Cham: Springer International Publishing, 2018.

[6] M. D. A. Muhajir, N. Prastiti Dan M. Koeshardianto, “Implementasi Chatbot Menggunakan Framework Langchain Berbasis Llm Gpt (Studi Kasus : Panduan Akademik Universitas Trunojoyo),” JATI (Jurnal Mahasiswa Teknik Informatika), Vol. 9, No. 2, Pp. 2151-2158, 2025.

[7] B.-S. Lee Dan B. R. Bryant, Automated Conversion From Requirements Documentation To An Object-Oriented Formal Specification Language, Madrid: SAC, 2014.

[8] S. Masri, H. I. Ashqar Dan M. Elhenawy, "Large Language Models (Llms) As Traffic Control Systems At Urban Intersections: A New Paradigm," 2024.

[9] N. A. Rakhmawati, Y. Awwab, A. C. Najib Dan A. Irsyad, "Ontology-Based Traffic Accident Information Extraction On Twitter In Indonesia," Inteligencia Artificial, Vol. 25, No. 70, Pp. 1-12, 2022.

[10] G. L. Pekerti, M. A. Yaqin Dan Suhartono, "Pencarian Model Proses Dalam Workflow Repository Berbasis Graph Database Menggunakan BPMN-Q," ILKOMNIKA: Journal Of Computer Science And Applied Informatics, Vol. 3, No. 1, Pp. 83-102, 2021.

[11] P. Kogler, A. Falkner, dan S. Sperl, "Reliable Generation of Formal Specifications using Large Language Models," Workshop Generative and Neurosymbolic AI in Software Engineering (GenSE 2024), Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn, pp. 1– 8, 2024. doi:10.18420/sw2024-ws_10.

[12] G. Granberry, W. Ahrendt, dan M. Johansson, "Specify What? A Case-Study using GPT-4 and Formal Methods For Specification Synthesis," The First AI for MATH Workshop at the 41st International Conference on Machine Learning, Vienna, Austria, 2024.

[13] M. N. Zitouni, A. A. Anda, S. Rajpal, D. Amyot, dan J. Mylopoulos, "Towards the LLM-Based Generation of Formal Specifications from Natural-Language Contracts: Early Experiments with SYMBOLEO," arXiv preprint arXiv:2411.15898, 2024. [Online]. Available: https://arxiv.org/abs/2411.15898.

[14] W. Murphy, N. Holzer, N. Koenig, L. Cui, R. Rothkopf, F. Qiao, dan M. Santolucito, "Guiding LLM Temporal Logic Generation with Explicit Separation of Data and Control," arXiv preprint arXiv:2406.07400, 2024. [Online]. Available: https://arxiv.org/abs/2406.07400.

Submitted : 04-06-2025 | Reviewed : 10-06-2025 | Accepted : 22-06-2025