ANALISIS AKURASI CHATGPT DALAM MENERJEMAHKAN ISTILAH TEKNIS TEKNIK ENERGI TERBARUKAN: PENDEKATAN HUMAN-IN-THE-LOOP

Nando Febriano Seva[1], Melsya Azka Aqila[2], Muhamad Rajib[3], Muhammad Kemal Pasya[4], Nasya Adzany Hidayanti[5], Mochamad Whilky Rizkyanfi[6]
nandoseva@student.upi.edu[1], melsyaazka@student.upi.edu[2], mjibrut@student.upi.edu[3], ruhiatkemalruhiat@student.upi.edu[4], nasyaadz@student.upi.edu[5], wilkysgm@upi.edu[6]
Universitas Pendidikan Indonesia

ABSTRAK

Penggunaan Large Language Model (LLM) seperti ChatGPT dalam penerjemahan mesin menawarkan efisiensi tinggi, namun menghadapi tantangan signifikan terkait akurasi terminologi domain-specific. Penelitian ini bertujuan mengukur akurasi ChatGPT dalam menerjemahkan istilah rekayasa teknik energi terbarukan dari bahasa Inggris ke bahasa Indonesia, serta mengevaluasi efektivitas pendekatan Human-in-the-Loop (HITL) dalam memvalidasi dan memperbaiki kesalahan terjemahan tersebut. Penelitian menggunakan metode campuran (mixed methods) dengan desain eksperimen evaluatif, menguji 48 istilah teknis yang diekstraksi dari glosarium standar IRENA dan IEA. Dataset dibagi ke dalam empat sub-domain: tenaga surya, tenaga angin, biomassa, dan penyimpanan energi. Pengujian dilaksanakan melalui empat fase HITL: zero-shot translation, validasi pakar, iterative prompt engineering, dan verifikasi akhir. Hasil penelitian menunjukkan bahwa pada kondisi zero-shot, akurasi dasar ChatGPT mencapai 83,33%. Sebanyak 39,6% istilah menunjukkan deviasi makna yang didominasi oleh Kesalahan Terminologi (68,4%), Kesalahan Konseptual (26,3%), dan Halusinasi Leksikal (5,3%), di mana domain penyimpanan energi mencatat tingkat kesalahan teknis tertinggi. Intervensi HITL oleh pakar terbukti efektif mengeliminasi seluruh kesalahan tersebut dan meningkatkan akurasi akhir secara absolut menjadi 100,00% melalui satu siklus iterasi koreksi. Disimpulkan bahwa ChatGPT layak digunakan sebagai alat terjemahan awal (first-pass translation), namun verifikasi pakar mutlak diperlukan guna mencegah risiko miskomunikasi pada dokumen rekayasa kritis. Penelitian ini turut menyoroti urgensi standardisasi glosarium nasional untuk sektor energi terbarukan.

Kata Kunci: Energi Terbarukan, Human-In-The-Loop, Penerjemahan Mesin, Terminologi Teknis.

PENDAHULUAN

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI), khususnya Large Language Model (LLM), telah membuka babak baru dalam dunia penerjemahan mesin. Salah satu model yang paling banyak diadopsi secara global adalah ChatGPT yang dikembangkan oleh OpenAI. Kehadiran ChatGPT telah menggeser paradigma penerjemahan dari pendekatan statistik dan berbasis aturan menuju pendekatan generatif berbasis konteks yang lebih fleksibel dan responsif terhadap kebutuhan pengguna [1].

Di sisi lain, sektor energi terbarukan (renewable energy) tengah mengalami pertumbuhan pesat, baik secara global maupun di Indonesia. Transisi energi dari bahan bakar fosil menuju sumber energi bersih seperti tenaga surya, angin, biomassa, dan hidrogen mendorong peningkatan kebutuhan komunikasi teknis lintas bahasa. Dokumen rekayasa, mulai dari spesifikasi teknis turbin angin, laporan efisiensi fotovoltaik, hingga panduan instalasi sistem penyimpanan energi, menuntut presisi terminologis yang mutlak. Pada konteks inilah keterbatasan LLM generalis dalam menangani terminologi domain-specific menjadi persoalan kritis, mengingat kesenjangan kualitas terjemahan yang dihasilkannya dapat berdampak langsung pada keselamatan dan keandalan sistem di lapangan [2].

Tantangan utama yang muncul adalah bahwa domain teknik energi terbarukan memiliki kosakata dan terminologi yang sangat spesifik, seperti photovoltaic efficiency,

grid parity, power purchase agreement (PPA), hingga islanding protection. Istilah-istilah tersebut tidak hanya memerlukan padanan linguistik yang tepat, tetapi juga pemahaman konteks teknis yang mendalam. Penelitian terdahulu menunjukkan bahwa meskipun ChatGPT mampu menghasilkan terjemahan yang fasih secara umum, model ini menghadapi kesulitan signifikan ketika berhadapan dengan terminologi domain-specific yang kompleks, terutama dalam konteks teknis dan rekayasa [3].

Kondisi ini semakin relevan di Indonesia, di mana adopsi teknologi energi terbarukan tengah dipercepat melalui berbagai kebijakan nasional. Kebutuhan akan penerjemahan dokumen teknis yang akurat antara bahasa Inggris dan bahasa Indonesia menjadi krusial, mengingat sebagian besar standar internasional, jurnal teknis, dan manual peralatan ditulis dalam bahasa Inggris. Kesalahan dalam penerjemahan istilah teknis dapat berakibat fatal, mulai dari miskomunikasi antara insinyur lokal dan tenaga ahli asing hingga kegagalan sistem yang berpotensi menimbulkan kerugian finansial dan risiko keselamatan. Oleh sebab itu, evaluasi akurasi ChatGPT dalam menerjemahkan istilah teknis energi terbarukan menjadi isu yang sangat mendesak untuk dikaji secara ilmiah.

Studi yang dilakukan oleh Ibrahim dkk. [4] secara eksplisit mencatat bahwa evaluasi ChatGPT di domain energi terbarukan yang telah dilakukan sebelumnya hanya berfokus pada informasi umum yang ditujukan untuk publik non-teknis, dan sama sekali tidak mengases konten teknis pada tingkat rekayasa (engineering-specific content). Kesenjangan ini menunjukkan bahwa belum ada penelitian yang secara sistematis menguji kemampuan ChatGPT dalam menangani terminologi teknis energi terbarukan, khususnya dalam konteks penerjemahan Inggris-Indonesia dengan melibatkan evaluasi pakar sebagai mekanisme penjaminan kualitas.

TINJAUAN PUSTAKA

Large Language Model dan Mekanisme ChatGPT

Large Language Model (LLM) adalah kelas model kecerdasan buatan yang dilatih pada korpus teks berskala masif menggunakan arsitektur *deep learning* berbasis *Transformer*. Konsep *Transformer* pertama kali diperkenalkan oleh Vaswani dkk. [11] melalui makalah landmark "*Attention Is All You Need*", yang mengajukan mekanisme *self-attention* sebagai alternatif yang lebih efisien dibandingkan arsitektur *recurrent neural network* (RNN) yang sebelumnya mendominasi pemrosesan bahasa alami (NLP). Berdasarkan fondasi arsitektur inilah keluarga model GPT (*Generative Pre-trained Transformer*) dikembangkan oleh OpenAI.

ChatGPT adalah antarmuka percakapan yang dibangun di atas fondasi model GPT (mulai dari GPT-3.5 hingga GPT-4o), yang keandalannya tidak sekadar bergantung pada skala parameter model, tetapi juga pada teknik pelatihan khusus yang disebut *Reinforcement Learning from Human Feedback* (RLHF). Menurut Hou dkk. [12], salah satu elemen kunci dalam desain *Transformer ChatGPT* adalah mekanisme *multi-head self-attention*, yang memungkinkan model memproses dan mengevaluasi beberapa posisi dalam urutan masukan secara paralel. Kemampuan ini memberikan *ChatGPT* keunggulan komparatif dalam memahami hubungan kontekstual antara kata dalam kalimat yang panjang dan kompleks.

Proses pelatihan *ChatGPT* berlangsung dalam tiga tahap utama. Pertama, model dilatih secara umum pada data teks dari internet dalam skala ratusan miliar token (*pre-training*). Kedua, model menjalani *Supervised Fine-Tuning* (SFT) menggunakan data demonstrasi yang dikurasi oleh pelabel manusia. Ketiga, model dioptimalkan lebih lanjut menggunakan algoritma *Proximal Policy Optimization* (PPO) yang mengacu pada skor dari model penghargaan yang telah dilatih berdasarkan preferensi manusia. Kombinasi ketiga tahap inilah yang menjadikan *ChatGPT* lebih responsif terhadap instruksi spesifik

dibandingkan model pendahulunya [13].

Meskipun arsitektur *Transformer* dan proses RLHF secara sinergis memungkinkan *ChatGPT* menghasilkan respons linguistik yang menyerupai manusia, kemampuan ini memiliki batasan yang nyata ketika model dihadapkan pada terminologi *domain-specific* yang tidak terwakili secara memadai dalam data pelatihan awalnya. Kajian empiris menunjukkan bahwa meskipun model LLM mampu bersaing dengan mesin NMT mutakhir untuk penerjemahan teks umum, terdapat kelemahan fundamental mengenai konsistensi terminologi dan penguasaan keahlian domain spesifik dalam keluaran terjemahan yang dihasilkannya [3].

Terminologi Teknik Energi Terbarukan dan Tantangan Penerjemahannya

Terminologi teknis merupakan satuan leksikal yang memiliki makna presisi dalam konteks bidang ilmu atau rekayasa tertentu. Berbeda dengan kosakata umum yang sering kali bersifat polisemik dan fleksibel, istilah teknis menuntut konsistensi makna mutlak dan ketepatan padanan lintas bahasa. Kesalahan minor dalam terjemahan istilah teknis dapat berakumulasi menjadi miskomunikasi yang fatal dalam implementasi rekayasa sistem di lapangan [14].

Teknik energi terbarukan merupakan salah satu domain rekayasa dengan perkembangan terminologi yang paling dinamis, sejalan dengan laju inovasi teknologi global di sektor ini. Berbeda dengan ranah medis atau hukum yang telah memiliki glosarium standar baku yang diterima luas, terminologi energi terbarukan terus beradaptasi mengikuti kemunculan sistem teknologi baru. Akibatnya, banyak entitas teknis di bidang ini yang belum memiliki padanan resmi atau konsisten dalam bahasa Indonesia. Kondisi ini memaksa penerjemah untuk memilih antara strategi penyerapan bahasa asing secara langsung (*borrowing*), penerjemahan deskriptif, atau penciptaan padanan kata baru yang sangat rentan memicu ambiguitas operasional.

Tantangan linguistik dalam ranah energi ini sebelumnya telah diinvestigasi oleh Ren dkk. [15] yang meneliti permasalahan penerjemahan istilah teknologi energi terbarukan antara bahasa Inggris dan Mandarin. Mereka menyoroti dua tantangan struktural utama: ketiadaan padanan langsung untuk istilah teknis yang sangat spesifik, dan kebutuhan adaptasi kontekstual mengingat sebuah istilah yang sama dapat memiliki makna yang jauh berbeda tergantung pada sub-domain aplikasinya. Temuan tersebut sangat relevan dengan konteks penerjemahan Inggris-Indonesia, di mana tekanan akurasi terminologis bahkan jauh lebih berat mengingat masih terbatasnya korpus teknis berbahasa Indonesia di domain rekayasa energi.

Tabel 1. Karakteristik Terminologi Teknis per Sub-domain Energi Terbarukan

Sub-domain	**Contoh Istilah (Bahasa Inggris)**	**Karakteristik dan Tantangan Penerjemahan (Bahasa Indonesia)**
Tenaga Surya	*Photovoltaic efficiency, Maximum Power Point Tracking* (MPPT)	Belum memiliki padanan baku yang seragam; umumnya diserap langsung sebagai akronim teknis atau diterjemahkan secara deskriptif kontekstual.
Tenaga Angin	*Tip speed ratio, Yaw control, Cut-in wind speed*	Mengandung makna mekanis yang sangat spesifik; terdapat risiko tinggi kehilangan makna parameter operasional jika diterjemahkan secara harfiah.
Biomassa	*Anaerobic digestion, Syngas,*	Melibatkan proses termokimia yang memerlukan pemahaman konteks rekayasa

	Torrefaction	proses; ketersediaan padanan dalam bahasa Indonesia masih sangat terbatas.
Penyimpanan Energi	*State of charge* (SoC), *Round-trip efficiency*, *Battery Management System* (BMS)	Didominasi oleh akronim teknis internasional yang wajib dipertahankan untuk standar operasional; terjemahan literal sering kali menciptakan kebingungan prosedural.

Keempat sub-domain tersebut merepresentasikan spektrum teknis yang sangat heterogen. Istilah-istilah dalam Tabel 1 tidak hanya sekadar nama komponen, melainkan merepresentasikan konsep perhitungan rekayasa yang memiliki implikasi operasional langsung di lapangan. Penerjemahan yang menyimpang berpotensi menyebabkan fatalitas dalam perancangan grid sistem, kesalahan kalibrasi instrumen, hingga kegagalan interpretasi pada laporan kinerja efisiensi suatu instalasi energi.

Dari perspektif NLP murni, rekayasa energi menghadirkan tantangan komputasi linguistik yang melampaui domain teks umum. Meegle [16] menegaskan bahwa kompleksitas terminologi di sektor energi membutuhkan arsitektur model dan prapemrosesan data pelatihan yang terspesialisasi. Hal ini sejalan dengan survei Chen dkk. [17] yang mengidentifikasi peningkatan signifikan dalam pengembangan model LLM khusus domain (*domain-specific LLM*) untuk sektor rekayasa energi, yang secara tidak langsung mengakui bahwa LLM generalis seperti *ChatGPT* masih rentan melakukan kesalahan konseptual ketika menangani dokumen teknis berisiko tinggi.

Konsep Human-in-the-Loop (HITL) dalam Validasi Terjemahan Teknis

Human-in-the-Loop (HITL) merupakan sebuah paradigma arsitektural dalam sistem kecerdasan buatan di mana intervensi manusia secara aktif diintegrasikan ke dalam satu atau lebih tahapan siklus kerja model AI. Tujuan utamanya adalah memastikan kualitas, meningkatkan akurasi, dan menjamin keandalan keluaran akhir. Dalam ranah penerjemahan mesin, mekanisme HITL umumnya diimplementasikan dalam alur kerja *post-editing*, yakni suatu proses verifikasi di mana seorang pakar domain (*Subject Matter Expert*) atau linguis profesional melakukan peninjauan, perbaikan, dan validasi terhadap keluaran sintesis AI sebelum digunakan dalam konteks operasional riil [8].

Bukti empiris paling komprehensif mengenai efikasi pendekatan HITL baru-baru ini ditunjukkan dalam studi komparatif penerjemahan medis oleh Brewster dkk. [18]. Studi tersebut membandingkan tingkat presisi antara sistem *ChatGPT* murni, penerjemah profesional manusia, dan pendekatan HITL (AI yang divalidasi pakar). Temuan riset tersebut menunjukkan bahwa sistem AI mandiri menghasilkan performa yang sangat fluktuatif dan rentan pada bahasa yang miskin korpus digital (*low-resource languages*). Sebaliknya, metodologi *Human-in-the-Loop* terbukti secara konsisten mampu menyamai bahkan melampaui kualitas terjemahan manusia profesional murni di seluruh parameter pengujian, terutama dalam mempertahankan presisi terminologi teknis.

Tabel 2. Perbandingan Metrik Evaluasi Otomatis dan Validasi Human-in-the-Loop (HITL)

Aspek Evaluasi	**Evaluasi AI Otomatis (Metrik BLEU/BERTScore)**	**Validasi Manusia (Human-in-the-Loop)**	**Keunggulan Signifikan HITL**
Akurasi Terminologi	Mengukur kemiripan leksikal semata; tidak memahami	Pakar menilai presisi semantik teknis dan	Menghasilkan terjemahan yang secara

	konteks keteknikan.	kelaziman istilah di dunia rekayasa.	engineering valid untuk domain rekayasa energi.
Deteksi Kesalahan Konseptual	Sangat rentan meloloskan kesalahan teknis selama tata bahasanya benar.	Manusia dapat mengidentifikasi secara langsung jika terdapat "halusinasi" yang melanggar hukum fisika/rekayasa.	Kesalahan fundamental operasional tidak dapat diotomasi dan wajib divalidasi langsung oleh ahli.
Konsistensi Glosarium	Parsial (hanya sebatas kemampuan exact-match).	Validasi menyeluruh berbasis komprehensi pengetahuan multidisiplin.	Mampu mengacu langsung pada standar atau nomenklatur industri kelistrikan lokal yang berlaku.
Justifikasi Operasional	Biaya dan komputasi rendah, sangat cepat untuk pemrosesan teks umum.	Membutuhkan waktu dan biaya validasi lebih tinggi, namun menihilkan kesalahan fatal.	Investasi waktu HITL sangat krusial dan wajib untuk validasi dokumen teknis berisiko keselamatan (safety-critical).

Analisis komparatif pada Tabel 2 mendemonstrasikan bahwa lapisan metodologi HITL memiliki keunggulan absolut yang tidak tergantikan oleh parameter matematis murni. Kompleksitas deteksi eror konseptual dan validasi kesesuaian glosarium operasional menjadikan peran validator ahli sebagai komponen penelitian yang esensial. Meskipun intervensi manusia berimplikasi pada peningkatan durasi evaluasi, hal ini merupakan syarat mutlak bagi standardisasi dokumen rekayasa energi terbarukan di mana presisi terjemahan berimplikasi langsung pada ketepatan desain, efisiensi investasi, dan keselamatan infrastruktur di lapangan.

METODE PENELITIAN

Penelitian ini menerapkan pendekatan metode campuran (mixed methods) dengan desain eksperimen evaluatif. Pendekatan ini dipilih untuk mengakomodasi kebutuhan pengukuran kuantitatif sekaligus pemahaman kualitatif yang mendalam. Secara kuantitatif, penelitian menghasilkan skor akurasi numerik dari terjemahan ChatGPT. Secara kualitatif, penelitian mendeskripsikan pola kesalahan terjemahan dan menganalisis konteks teknis di

balik kegagalan terminologi tersebut.

Penggunaan mixed methods ditujukan untuk mencapai validitas komprehensif. Saraswati dan Devi [19] menegaskan bahwa paradigma ini memberikan perspektif yang lebih holistik terhadap permasalahan yang multifaset. Desain ini juga sejalan dengan preseden pada domain evaluasi AI, di mana triangulasi data kuantitatif dan kualitatif terbukti menghasilkan temuan yang lebih kaya [20].

Secara operasional, komponen kuantitatif mencakup perhitungan skor akurasi agregat per fase HITL, distribusi frekuensi kesalahan, dan pengukuran persentase peningkatan akurasi. Sementara itu, komponen kualitatif berfokus pada analisis tematik pola kesalahan, kategorisasi jenis eror, dan deskripsi naratif konteks kegagalan AI. Desain eksperimen evaluatif memungkinkan peneliti untuk tidak sekadar mengobservasi, tetapi secara aktif mengukur dampak kausal dari intervensi prompt engineering dan validasi pakar terhadap kualitas keluaran akhir.

HASIL DAN PEMBAHASAN

Akurasi Terjemahan ChatGPT pada Fase 1 (Zero-Shot Translation)

Pada Fase 1, *ChatGPT* diajukan 48 istilah teknis energi terbarukan menggunakan *prompt* minimal tanpa informasi domain tambahan. Seluruh terjemahan yang dihasilkan kemudian dinilai oleh pakar menggunakan rubrik evaluasi tiga tingkat (skor 1-3). Hasil pengukuran menunjukkan bahwa dari total skor maksimum 144 poin, *ChatGPT* berhasil memperoleh 120 poin, menghasilkan nilai akurasi *baseline* (A_1) sebesar 83,33%.

Distribusi skor secara keseluruhan menunjukkan bahwa 29 istilah (60,4%) mendapatkan skor 3 (Akurat), 14 istilah (29,2%) mendapatkan skor 2 (Kurang Akurat), dan 5 istilah (10,4%) mendapatkan skor 1 (Tidak Akurat). Rincian distribusi skor disajikan pada Tabel 6.

Tabel 6. Distribusi Skor Akurasi Terjemahan ChatGPT pada Fase 1 (Zero-Shot)

Skor	Kategori	Jumlah Persentase (%)	Interpretasi
3	Akurat	29 (60,4%)	Terjemahan tepat secara linguistik dan teknis.
2	Kurang Akurat	14 (29,2%)	Terjemahan harfiah benar namun tidak lazim dalam rekayasa.
1	Tidak Akurat	5 (10,4%)	Terjemahan salah konsep, menyesatkan, atau halusinasi.
	Total: 48 (100,0%)		

Data distribusi skor tersebut mengindikasikan bahwa *ChatGPT* memiliki kemampuan dasar yang cukup baik dalam menerjemahkan terminologi teknis energi terbarukan secara umum, dengan mayoritas istilah (60,4%) berhasil diterjemahkan dengan akurasi penuh. Namun demikian, hampir 40% istilah menunjukkan permasalahan pada tingkat yang bervariasi. Proporsi ini signifikan mengingat sifat kritis terminologi teknis dalam konteks rekayasa, di mana ketidaktepatan terjemahan dapat berdampak langsung pada implementasi sistem.

Temuan ini selaras dengan hasil penelitian Moslem dkk. [7] yang menemukan kesenjangan kualitas yang signifikan antara LLM generalis dan model terjemahan *domain-specific*, terutama untuk terminologi teknis tingkat tinggi. Nilai A_1 sebesar 83,33% juga

konsisten dengan temuan Albuhairy dan Algaraady [3] yang menyimpulkan bahwa meskipun *ChatGPT* mampu menghasilkan terjemahan yang fasih secara umum, model ini menghadapi tantangan signifikan pada terminologi *domain-specific* yang kompleks, terutama istilah yang memerlukan pemahaman konteks operasional rekayasa.

Analisis Akurasi per Sub-domain

Analisis per sub-domain mengungkapkan variasi akurasi yang bermakna di antara keempat area teknis yang diuji. Tabel 7 menyajikan rekapitulasi lengkap akurasi pada Fase 1 dan Fase 4 untuk setiap sub-domain.

Tabel 7. Rekapitulasi Akurasi Terjemahan ChatGPT per Sub-domain (Fase 1 dan Fase 4)

Sub-domain	Jumlah Istilah	Σ Skor Fase 1	A_1 (%)	Σ Skor Fase 4	A_2 (%)
Tenaga Surya	12	31	86,11	36	100,00
Tenaga Angin	12	29	80,56	36	100,00
Biomassa	12	33	91,67	36	100,00
Penyimpanan Energi	12	27	75,00	36	100,00
Total / Rata-rata	**48**	**120**	**83,33**	**144**	**100,00**

Dari Tabel 7 terlihat bahwa pada Fase 1, sub-domain Biomassa mencatat akurasi tertinggi sebesar 91,67%, diikuti Tenaga Surya (86,11%), Tenaga Angin (80,56%), dan Penyimpanan Energi sebagai sub-domain dengan akurasi terendah sebesar 75,00%.

Sub-domain Tenaga Surya (A_1 = 86,11%)

Dari 12 istilah teknis tenaga surya yang diuji, 8 istilah (66,7%) mendapatkan skor penuh 3. Empat istilah menunjukkan permasalahan: *Ingress Protection* (IP) *rating* mendapatkan skor 1 karena *ChatGPT* menghasilkan terjemahan "Tingkat perlindungan masuk (IP)" yang secara terminologi salah. Padanan baku dalam standar kelistrikan adalah "Indeks Proteksi". Tiga istilah lainnya mendapatkan skor 2, yaitu *Maximum Power Point Tracking* (MPPT) yang diterjemahkan dengan penambahan kata "Pelacakan" yang tidak lazim (padanan industri: "MPPT"), *Fill factor* yang dikacaukan dengan konteks pengisian baterai ("Faktor pengisian"), dan *Grid-tie inverter* yang menghasilkan terjemahan deskriptif "Inverter terhubung jaringan" alih-alih padanan standar industri "Inverter *on-grid*".

Sub-domain Tenaga Angin (A_1 = 80,56%)

Sub-domain tenaga angin mencatat 6 istilah bermasalah dari 12 istilah yang diuji. Seluruh 6 kesalahan tersebut dikategorikan sebagai Kesalahan Terminologi. Pola yang dominan adalah kecenderungan *ChatGPT* menambahkan kata penjelas dalam kurung yang tidak diperlukan dan tidak lazim dalam komunitas rekayasa, seperti "Kontrol *yaw* (arah)", "Sudut *pitch* (sudut bilah)", dan "Kecepatan angin minimum (mulai beroperasi)". Selain itu, istilah *Nacelle* diterjemahkan sebagai "Nacelle (rumah mesin turbin)" padahal serapan resmi sesuai kaidah teknik adalah "Nasel". Kasus paling kritis adalah *Wind shear* yang diterjemahkan sebagai "Geser angin" (skor 1), menggunakan bentuk kata kerja alih-alih nomina "Geseran angin" yang merupakan padanan teknis yang tepat.

Sub-domain Biomassa (A_1 = 91,67%)

Biomassa mencatat akurasi tertinggi dengan hanya 2 istilah bermasalah. Mayoritas istilah kimia-teknis seperti *torrefaction*, *pyrolysis*, *gasification*, dan *transesterification* berhasil diterjemahkan dengan tepat melalui pola serapan morfologis yang konsisten. Kegagalan terjadi pada *Anaerobic digestion* (skor 1) yang diterjemahkan sebagai "Pencernaan anaerob", menggunakan kosakata medis/biologis yang tidak sesuai konteks

bioteknologi rekayasa; padanan yang tepat adalah "Digesti anaerobik". Istilah *Co-firing* mendapatkan skor 2 karena terjemahan "Pembakaran bersama" menghilangkan nuansa industri yang terkandung dalam konsep pembakaran pendamping bahan bakar fosil dengan biomassa secara simultan.

Sub-domain Penyimpanan Energi (A_1 = 75,00%)

Sub-domain Penyimpanan Energi mencatat akurasi terendah dengan 7 dari 12 istilah (58,3%) menunjukkan permasalahan. Temuan ini mengindikasikan bahwa terminologi sistem penyimpanan energi merupakan area paling menantang bagi *ChatGPT*. Empat istilah mendapatkan skor 1 atau 2 karena kesalahan konseptual: (a) *State of Charge* (SoC) diterjemahkan sebagai "Tingkat pengisian" yang ambigu dengan laju pengisian daya; (b) *Round-trip efficiency* diterjemahkan sebagai "Efisiensi siklus penuh" yang mengaburkan sifat bolak-balik dari siklus penyimpanan-pelepasan; (c) *Depth of Discharge* (DoD) menghasilkan "Kedalaman pelepasan" yang mengaburkan makna elektronika baterai; (d) *Thermal runaway* diterjemahkan sebagai "Pelarian termal", yang merupakan terjemahan literal tanpa pemahaman konteks kegagalan termis dan menjadi satu-satunya kasus Halusinasi Leksikal dalam dataset ini.

Analisis Pola Kesalahan Terjemahan

Dari 48 istilah yang diuji, terdapat 19 istilah (39,6%) yang menunjukkan permasalahan pada Fase 1. Analisis kategoris terhadap kesalahan tersebut menghasilkan tiga tipologi utama sesuai kerangka MQM [28] sebagaimana disajikan pada Tabel 8.

Tabel 8. Distribusi Kategori Kesalahan Terjemahan ChatGPT pada Fase 1

Kategori Kesalahan	Jumlah Kasus	Persentase (%)
Kesalahan Terminologi	13	68,4
Kesalahan Konseptual	5	26,3
Halusinasi Leksikal	1	5,3
Total	**19**	**100,0**

Kesalahan Terminologi (68,4% dari Total Kesalahan)

Kesalahan Terminologi merupakan tipologi yang paling dominan, mencakup 13 dari 19 kasus kesalahan (68,4%). Pola ini terjadi ketika *ChatGPT* menghasilkan padanan kata yang secara harfiah dapat diterima namun tidak sesuai dengan glosarium standar yang berlaku dalam komunitas teknik energi internasional. Akar permasalahan tipologi ini adalah ketiadaan representasi glosarium *domain-specific* yang memadai dalam data pelatihan *ChatGPT*, sehingga model cenderung menggunakan strategi terjemahan komposisional (menerjemahkan komponen kata per kata) alih-alih mengakses padanan terminologi yang telah ditetapkan secara konvensional. Contoh representatif adalah terjemahan "Pelacakan Titik Daya Maksimum" untuk MPPT, yang secara harfiah dapat dipahami namun tidak pernah digunakan dalam dokumen teknis energi surya maupun standar IEC yang relevan.

Kesalahan Konseptual (26,3% dari Total Kesalahan)

Kesalahan Konseptual mencakup 5 kasus (26,3%) dan merupakan tipologi yang paling berbahaya dari perspektif rekayasa, karena terjemahan yang dihasilkan tidak hanya tidak lazim, tetapi secara aktif mengaburkan atau mengubah makna teknis asli istilah. Kasus paling ilustratif adalah "*Anaerobic digestion*" yang diterjemahkan sebagai "Pencernaan anaerob", meminjam terminologi dari domain medis/biologi yang secara konteks teknis berbeda signifikan. Dalam rekayasa bioenergi, istilah ini merujuk pada proses biokimia terstruktur dalam reaktor anaerob, bukan proses fisiologis pencernaan. Kesalahan serupa terjadi pada "*Depth of Discharge*" dan "*Round-trip efficiency*" di mana interpretasi

ChatGPT menunjukkan kurangnya pemahaman terhadap dinamika siklus penyimpanan energi elektrokimia.

Halusinasi Leksikal (5,3% dari Total Kesalahan)

Halusinasi Leksikal teridentifikasi pada satu kasus, yaitu istilah *Thermal runaway* yang diterjemahkan sebagai "Pelarian termal". Meskipun hanya satu kasus, tipologi ini memiliki implikasi serius karena terjemahan yang dihasilkan tidak hanya salah, tetapi berpotensi menyesatkan secara kritis. *Thermal runaway* dalam rekayasa baterai litium-ion adalah fenomena kegagalan termis kaskade yang dapat mengakibatkan ledakan atau kebakaran; mengalihbahasakan konsep ini sebagai "pelarian" menghilangkan seluruh konteks darurat dan risiko keselamatan yang terkandung dalam istilah aslinya. Temuan ini selaras dengan peringatan Albuhairy dan Algaraady [3] bahwa *ChatGPT* dapat menghasilkan terjemahan yang tidak akurat tanpa optimasi konteks domain yang memadai.

Kontribusi Pendekatan Human-in-the-Loop terhadap Peningkatan Akurasi

Fase 3 dan Fase 4 dari siklus HITL secara bersama-sama mendemonstrasikan kontribusi yang signifikan dari intervensi manusia terhadap kualitas terjemahan teknis *ChatGPT*. Tabel 9 menyajikan perbandingan langsung antara terjemahan Fase 1 dan Fase 4 pada istilah-istilah yang sebelumnya bermasalah.

Tabel 9. Perbandingan Terjemahan Fase 1 vs. Fase 4 pada Istilah Bermasalah

Istilah Asli	Terjemahan Fase 1	Skor	Terjemahan Fase 4 (HITL)	Skor
Ingress Protection (IP) rating	Tingkat perlindungan masuk (IP)	1	Indeks Proteksi (IP) / Rating IP	3
Wind shear	Geser angin	1	Geseran angin	3
Anaerobic digestion	Pencernaan anaerob	1	Digesti anaerobik	3
Depth of Discharge (DoD)	Kedalaman pelepasan (DoD)	1	Kedalaman pengosongan daya (DoD)	3
Thermal runaway	Pelarian termal	1	Kegagalan termal	3
Fill factor	Faktor pengisian	2	Faktor isi	3
Grid-tie inverter	Inverter terhubung jaringan	2	Inverter on-grid / tersambung jala-jala	3
Nacelle	Nacelle (rumah mesin turbin)	2	Nasel	3
Round-trip efficiency	Efisiensi siklus penuh	2	Efisiensi siklus bolak-balik	3
C-rate	Laju C	2	C-rate	3

Dari Tabel 9 terlihat bahwa seluruh 19 istilah yang bermasalah pada Fase 1 berhasil diperbaiki melalui intervensi *iterative prompt engineering* pada Fase 3 dan diverifikasi mencapai skor 3 (Akurat) pada Fase 4. Secara kuantitatif, nilai akurasi meningkat dari A_1 = 83,33% pada Fase 1 menjadi A_2 = 100,00% pada Fase 4, dengan persentase peningkatan ΔA = 16,67 poin persentase.

Peningkatan 16,67 poin persentase ini memiliki makna praktis yang substansial. Dalam konteks dataset 48 istilah dengan skor maksimum 144, selisih 24 poin skor merepresentasikan 19 istilah yang berhasil ditransformasi dari terjemahan yang salah atau tidak lazim menjadi terjemahan yang akurat secara teknis dan diterima oleh komunitas rekayasa. Dengan kata lain, pendekatan HITL berhasil mengeliminasi 100% kesalahan terjemahan yang teridentifikasi pada kondisi *zero-shot*.

Temuan ini secara empiris mengonfirmasi proposisi teoritis yang diajukan oleh Brewster dkk. [18], bahwa terjemahan *Human-in-the-Loop* secara konsisten mencapai hasil yang sebanding atau lebih baik dari terjemahan profesional murni. Dalam konteks penelitian ini, HITL tidak sekadar meningkatkan kualitas, melainkan mentransformasi keluaran dari tidak memadai menjadi memadai untuk penggunaan dalam dokumen rekayasa teknis.

Pembahasan Implikasi Temuan

Secara keseluruhan, temuan penelitian ini memiliki tiga implikasi utama yang relevan bagi komunitas akademis maupun praktisi di bidang teknik energi terbarukan dan teknologi terjemahan.

Pertama, implikasi terhadap penggunaan *ChatGPT* sebagai alat terjemahan teknis. Nilai akurasi *zero-shot* sebesar 83,33% menunjukkan bahwa *ChatGPT* memiliki potensi yang memadai sebagai alat bantu terjemahan awal (*first-pass translation*) untuk dokumen teknis energi terbarukan, namun tidak dapat diandalkan sebagai satu-satunya sumber tanpa validasi domain. Penggunaan *ChatGPT* tanpa mekanisme verifikasi teknis berisiko menghasilkan dokumen yang mengandung terminologi yang tidak lazim (29,2% istilah) atau bahkan menyesatkan secara konseptual (10,4% istilah). Risiko ini paling tinggi pada domain Penyimpanan Energi (A_1 = 75%) dan paling rendah pada domain Biomassa (A_1 = 91,67%).

Kedua, implikasi terhadap desain sistem terjemahan berbasis AI untuk domain teknis. Keberhasilan HITL dalam meningkatkan akurasi dari 83,33% menjadi 100% melalui satu siklus iterasi menunjukkan bahwa arsitektur sistem terjemahan teknis yang optimal adalah model hibrid: AI sebagai generator terjemahan awal, dan pakar domain sebagai validator serta penyedia konteks korektif. Desain ini lebih efisien dibandingkan terjemahan manusia murni sekaligus lebih andal dibandingkan AI murni.

Ketiga, implikasi terhadap pengembangan glosarium teknis bahasa Indonesia. Pola kesalahan yang didokumentasikan dalam penelitian ini, terutama dominasi Kesalahan Terminologi (68,4%), mengindikasikan bahwa permasalahan mendasar bukanlah semata-mata keterbatasan *ChatGPT*, melainkan juga ketiadaan glosarium teknis energi terbarukan yang terstandarisasi dalam bahasa Indonesia. Pengembangan glosarium resmi yang mengacu pada standar IRENA [21] dan IEA [22] dalam bahasa Indonesia akan secara langsung meningkatkan kualitas terjemahan AI maupun manusia di domain ini.

KESIMPULAN

Penelitian ini berhasil menjawab dua tujuan yang ditetapkan, yaitu mengukur akurasi terjemahan ChatGPT pada terminologi teknis rekayasa energi terbarukan dan mengevaluasi kontribusi pendekatan Human-in-the-Loop (HITL) dalam meningkatkan kualitas terjemahan tersebut. Berdasarkan analisis terhadap 48 istilah teknis yang dikompilasi dari glosarium IRENA dan IEA dalam empat fase siklus HITL, penelitian ini menghasilkan tiga kesimpulan utama:

1. Akurasi Zero-Shot Belum Memadai untuk Standar Rekayasa. ChatGPT menunjukkan akurasi yang cukup baik namun belum memadai untuk penggunaan mandiri dalam penerjemahan dokumen teknis. Pada kondisi zero-shot (Fase 1), ChatGPT mencatat nilai akurasi baseline (A1) sebesar 83,33%. Meskipun 60,4% istilah berhasil diterjemahkan

dengan tepat, sebanyak 39,6% istilah menunjukkan permasalahan berupa Kesalahan Terminologi (68,4%), Kesalahan Konseptual (26,3%), dan Halusinasi Leksikal (5,3%). Sub-domain Penyimpanan Energi merupakan area paling kritis dengan proporsi kesalahan konseptual tertinggi.

2. Efektivitas Intervensi Human-in-the-Loop. Pendekatan HITL terbukti secara empiris mampu mengeliminasi seluruh kesalahan terminologi yang teridentifikasi pada kondisi awal. Melalui siklus intervensi validasi pakar dan iterative prompt engineering, akurasi terjemahan meningkat secara signifikan menjadi 100,00% (A2), dengan persentase peningkatan (ΔA) sebesar 16,67 poin persentase. Hal ini menegaskan bahwa HITL tidak sekadar meningkatkan kualitas, melainkan mentransformasi keluaran menjadi memadai secara operasional untuk kebutuhan rekayasa di lapangan.
3. Keterbatasan Sistemis Glosarium Nasional. Pola kesalahan dominan menunjukkan adanya keterbatasan sistemis yang bersumber dari ketiadaan representasi glosarium teknis domain energi terbarukan dalam bahasa Indonesia. Dominasi Kesalahan Terminologi mengindikasikan bahwa ChatGPT cenderung menggunakan strategi terjemahan komposisional alih-alih mengakses padanan terminologi konvensional. Permasalahan akurasi ini membutuhkan solusi komprehensif yang melampaui optimasi model AI semata.

Saran

Berdasarkan temuan dan keterbatasan penelitian ini, terdapat empat rekomendasi yang ditujukan kepada pihak-pihak terkait:

Bagi Praktisi Teknik: ChatGPT dapat dimanfaatkan sebagai alat bantu terjemahan awal (first-pass translation), dengan catatan wajib disertai mekanisme verifikasi oleh pakar domain sebelum dokumen digunakan secara operasional. Tim teknik disarankan menyusun glosarium kerja internal yang mengacu pada standar IRENA dan IEA sebagai panduan prompt engineering yang konsisten.

Bagi Peneliti Lanjutan: Diperlukan studi komparatif yang membandingkan akurasi ChatGPT dengan model LLM lain (seperti Gemini atau Claude) pada dataset yang sama. Selain itu, perluasan dataset dan pengembangan metodologi yang mengintegrasikan metrik evaluasi otomatis (BLEU, BERTScore) dengan penilaian HITL secara paralel sangat direkomendasikan.

Bagi Pengembang Sistem AI: Arsitektur sistem terjemahan teknis yang optimal adalah model hibrid yang mengintegrasikan LLM generalis dengan basis pengetahuan terminologi domain-specific. Desain instrumen evaluasi berbasis rubrik MQM yang terstandarisasi perlu menjadi komponen wajib dalam pengembangannya.

Bagi Pemangku Kebijakan: Disarankan kepada Badan Pengembangan dan Pembinaan Bahasa (Badan Bahasa) bersama Kementerian Energi dan Sumber Daya Mineral (ESDM) untuk memprioritaskan penyusunan Glosarium Nasional Teknik Energi Terbarukan. Glosarium ini akan menjadi referensi baku bagi insinyur, akademisi, dan validator AI di seluruh Indonesia.

DAFTAR PUSTAKA

Z. Jiang, Q. Lv, Z. Zhang, and L. Lei, "Convergences and divergences between automatic assessment and human evaluation: Insights from comparing ChatGPT-generated translation and neural machine translation," arXiv, 2024. https://doi.org/10.48550/arXiv.2401.05176

J. Zheng, H. Hong, F. Liu, X. Wang, J. Su, Y. Liang, and S. Wu, "Fine-tuning large language models for domain-specific machine translation," arXiv, 2024. https://doi.org/10.48550/arXiv.2402.15061

S. Albuhairy and J. Algaraady, "Exploring ChatGPT's potential for augmenting post-editing in machine translation across multiple domains: Challenges and opportunities," Frontiers in

Artificial Intelligence, vol. 8, 2025. https://doi.org/10.3389/frai.2025.1526293
M. S. Ibrahim, A. J. Aljaaf, M. Al-khafajiy, A. A. Nafea, and N. S. Sani, "Systematic evaluation of ChatGPT performance in providing renewable energy information," PeerJ Computer Science, vol. 11, p. e3295, 2025. https://doi.org/10.7717/peerj-cs.3295
X. Li, Y. Zhang, and H. Chen, "GPT-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels," arXiv, 2024. https://arxiv.org/pdf/2407.03658
L. Besacier, M. Alam, M. Galle, V. Nikoulina, and A. Antonis, "On the evaluation of machine translation for terminology consistency," arXiv, 2021. https://arxiv.org/pdf/2106.11891
Y. Moslem, R. Haque, and A. Way, "Domain-specific translation with open-source large language models: Resource-oriented analysis," arXiv, 2024. https://doi.org/10.48550/arXiv.2412.05862
Machine Translate, "Human-in-the-loop," [Online]. Available: https://machinetranslate.org/human-in-the-loop
Cobbai, "Quality at scale: Best practices for MT and human-in-the-loop translation workflows," Nov. 24, 2025. [Online]. Available: https://cobbai.com/blog/translation-quality-support
Contentech, "Human-in-the-loop vs fully automated translation: Which is better for your business?" Feb. 26, 2026. [Online]. Available: https://contentech.com/human-in-the-loop-vs-fully-automated-translation
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv, 2017. https://doi.org/10.48550/arXiv.1706.03762
Z. Hou et al., "Systematic exploration and in-depth analysis of ChatGPT architectures progression," Artificial Intelligence Review, vol. 57, 2024. https://doi.org/10.1007/s10462-024-10832-0
N. Lambert, "Reinforcement learning from human feedback," arXiv, 2025. https://doi.org/10.48550/arXiv.2504.12501
P. Naveen and P. Trojovsky, "Overview and challenges of machine translation for contextually appropriate translations," iScience, vol. 27, no. 10, p. 110878, 2024. https://doi.org/10.1016/j.isci.2024.110878
J. Ren, A. Wang, and L. Su, "Translation and context adaptability study of renewable energy technology terms in English and Chinese," International Journal of Emerging Electric Power Systems, vol. 26, no. 6, pp. 1017-1031, 2025. https://doi.org/10.1515/ijeeps-2024-0395
Meegle, "Natural language processing for energy," 2024. [Online]. Available: https://www.meegle.com/en_us/topics/natural-language-processing/natural-language-processing-for-energy
Y. Chen et al., "A comprehensive survey of LLMs for sustainable and renewable energy systems," MDPI Information, vol. 17, no. 3, p. 271, 2026. https://doi.org/10.3390/info17030271
R. C. Brewster, G. Tse, A. L. Fan et al., "Evaluating human-in-the-loop strategies for artificial intelligence-enabled translation of patient discharge instructions: A multidisciplinary analysis," npj Digital Medicine, vol. 8, p. 629, 2025. https://doi.org/10.1038/s41746-025-02055-6
P. Saraswati and A. Devi, "Mixed methods-research methodology: An overview," Mathews Journal of Nursing, vol. 5, no. 4, Art. 24, 2023.
X. Quan and Y. Sun, "Artificial intelligence-powered evaluation model for English translation education in university: Combining quantitative and qualitative methods," Scientific Reports, 2026. https://doi.org/10.1038/s41598-026-46314-2
International Renewable Energy Agency, "Renewable energy statistics and glossary," 2023. [Online]. Available: https://www.irena.org/Statistics
International Energy Agency, "IEA glossary," 2023. [Online]. Available: https://www.iea.org/glossary
Sugiyono, Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta, 2019.
K. A. Yuksel, A. Gunduz, A. B. Anees, and H. Sawaf, "Efficient machine translation corpus generation: Integrating human-in-the-loop post-editing with large language models," arXiv, 2025. https://doi.org/10.48550/arXiv.2502.12755

M. Yamada, "Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability," in Proc. Machine Translation Summit XIX, Vol. 2: Users Track, 2023, pp. 195-204.
M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey, "Experts, errors, and context: A large-scale study of human evaluation for machine translation," arXiv, 2021. https://doi.org/10.48550/arXiv.2104.14478
R. B. Johnson, A. J. Onwuegbuzie, and L. A. Turner, "Toward a definition of mixed methods research," Journal of Mixed Methods Research, vol. 1, no. 2, pp. 112-133, 2007. https://doi.org/10.1177/1558689806298224
A. Lommel, H. Uszkoreit, and A. Burchardt, "Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics," Tradumàtica, no. 12, pp. 455-463, 2014. https://doi.org/10.5565/rev/tradumatica.77