The Use of ChatGpt as an Arabic Translation Tool: A Study of BLEU Metrics Quality and Human Assessment
Pemanfaatan ChatGpt sebagai Alat Terjemahan Bahasa Arab: Kajian Kualitas Metrik BLEU dan Penilaian Manusia

Rahma Auliyaa Izzati[1)], Farikh Marzuki Ammar*[,2)]

[1)] Program Studi Pendidikan Bahasa Arab,Universitas Muhammadiyah Sidoarjo
[2)] ProgramSrudi Pendidikan Bahasa Arab, Univeesitas Muhammadiya Sidoarjo
* Email Penulis Korespondensi : farikh1@umsida.ac.id[2)]

Abstract. *This study aims to analyze the quality of ChatGPT translations from Arabic to Indonesian using a combined approach of automated evaluation based on the BLEU metric and human evaluation. The method used was descriptive quantitative analysis, involving 30 students as evaluators who assessed accuracy, acceptability and readability. The results show an average BLEU score of 0.4123, indicating moderate alignment with fluctuating variations. Human evaluation revealed that translation quality falls into the high category for accuracy and readability and the moderate category for acceptability. Furthermore the results indicate that BLEU scores and human evaluations do not always yield similar assessments because the two methods measure different aspects of translation quality. Therefore translation quality evaluation should be conducted in a complementary manner. These findings suggest that ChatGPT has the potential to serve as a translation aid but still requires human evaluation to achieve optimal results.*
Keywords - *ChatGPT, Arabic translation, BLEU*

Abstrak. *Penelitian ini bertujuan untuk menganalisis kualitas terjemahan ChatGPT dari Bahasa Arab ke Bahasa Indonesia dengan menggunakan pendekatan gabungan antara evaluasi otomatis berbasis metrik BLEU dan evaluasi manusia. Metode yang digunakan adalah kuantitatif deskriptif dengan melibatkan 30 mahasiswa sebagai evaluator pada aspek keakuratan, keberterimaan dan keterbacaan. Hasil penelitian menunjukkan skor rata-rata BLEU sebesar 0,4123 yang mengindikasikan kesesuaian moderat dengan variasi yang fluktuatif. Evaluasi manusia menunjukkan kualitas terjemahan berada pada kategori tinggi untuk keakuratan dan keterbacaan serta sedang untuk keberterimaan. Selain itu hasil penelitian menunjukkan bahwa skor BLEU dan evaluasi manusia tidak selalu memberikan penilaian yang serupa karena kedua metode mengukur aspek kualitas terjemahan yang berbeda. Oleh karena itu evaluasi kualitas terjemahan perlu dilakukan secara komplementer. Temuan ini menunjukkan bahwa ChatGPT berpotensi sebagai alat bantu penerjemahan namun tetap memerlukan evaluasi manusia untuk hasil yang optimal.*
Kata kunci – *ChatGPT, penerjemahan Bahasa Arab, BLEU*

I. PENDAHULUAN

Pembelajaran Bahasa Arab di Indonesia menghadapi tantangan yang cukup kompleks sebagai bahasa yang memiliki sistem morfologi (*sarf*), sintaksis (*nahwu*) dan semantik (*dalalah*) yang rumit[1] karena bahasa arab menuntut pemahaman mendalam terhadap struktur dan konteks makna. Kompleksitas ini menjadikan pembelajaran Bahasa Arab lebih menantang dibandingkan dengan bahasa asing lainnya terutama karena perbedaan mendasar antara struktur Bahasa Arab dan Bahasa Indonesia[2]. Dalam struktur Bahasa Arab satu kata dasar (*fi'il*) dapat berubah bentuk menjadi puluhan derivasi morfologis, sementara Bahasa Indonesia cenderung lebih sederhana dalam pembentukan kata[3]. Perbedaan mendasar inilah yang menjadi salah satu hambatan utama yang harus dihadapi pelajar di indonesia dalam menguasai keterampilan membaca, menulis serta menerjemahkan teks bahasa arab.

Selain itu, Bahasa Arab mempunyai keragaman linguistik yang relatif kompleks yang mencakup bahasa arab klasik (*Classical Arabic)*, Bahasa arab modern (*Modern Standart Arabic)*[4] serta berbagai dialek lokal yang mirip seperti bahasa arab mesir, bahasa arab suriah dan bahasa arab teluk atau *Gulf Arabic*[5]. Setiap dialek ini memiliki perbedaan yang jelas pada tingkat fonetik, morfologi dan tata bahasa[6]. Perbedaan-perbedaan ini meningkatkan tingkat kesulitan bagi pembelajar sehingga memahami dan menerjemahkan teks arab dengan akurat menjadi lebih rumit. Dalam konteks akademis sebagian besar teks yang dipelajari di Indonesia berasal dari bahasa Arab klasik (*fushah*) yang digunakan dalam buku-buku turats, sementara media modern lebih banyak menggunakan *MSA* yang

lebih ringkas dan komunikatif. Ketidaksamaan karakteristik ini membuat mahasiswa sering kali menghadapi kebingungan dalam memahami makna sebenarnya dari suatu teks.

Kesulitan utama dalam pembelajaran Bahasa Arab di Indonesia tidak hanya terletak pada aspek gramatikal[7] tetapi juga pada penerjemahan makna idiomatik dan konteks budaya[8]. Penerjemahan Bahasa Arab ke Bahasa Indonesia menuntut kemampuan memahami makna tersirat agar hasil terjemahan tetap alami, komunikatif dan sesuai konteks. Menurut Huda, banyak kesalahan penerjemahan di kalangan mahasiswa terjadi karena kurangnya pemahaman terhadap unsur budaya dan idiomatik yang terkandung dalam teks sumber. Misalnya ungkapan idiomatik yang terdapat dalam kitab Syarah-syarah Ad Diwan Imam Syafi'i yang berbunyi " 'A'ma qalbi " *(Buta Hati*) jika diterjemahkan secara harfiah ke Bahasa Indonesia menjadi tidak alami sementara secara kontekstual seharusnya diartikan "Tidak mau menerima kebenaran"[9]. Hal ini menunjukkan bahwa penguasaan linguistik semata belum cukup untuk menghasilkan terjemahan yang baik akan tetapi dibutuhkan juga pemahaman pragmatik dan kultural yang mendalam.

Tantangan nyata yang dihadapi mahasiswa dalam menerjemahkan teks Arab klasik muncul ketika mereka berhadapan dengan istilah khas daerah Arab tertentu yang menuntut pemahaman terhadap konteks sosial dan budaya penuturnya. Kesalahan struktural dan semantik sering kali terjadi karena penerjemah tidak mempertimbangkan aspek pragmatik yakni faktor-faktor yang memengaruhi pilihan bahasa dalam konteks interaksi sosial. Sejalan dengan pandangan Suyono, penggunaan bahasa tidak hanya bergantung pada penguasaan kaidah gramatikal tetapi juga pada pemahaman terhadap kaidah sosiokultural serta konteks pemakaian bahasa. Dengan demikian, sebagaimana dikemukakan Verhaar pragmatik berperan penting dalam memahami bagaimana struktur bahasa berfungsi sebagai alat komunikasi antara penutur dan pendengar serta keterkaitannya dengan hal-hal di luar bahasa itu sendiri [10]. Misalnya kalimat majemuk dalam teks klasik sering kali memiliki pola inversi atau susunan terbalik yang tidak umum dalam Bahasa Indonesia sehingga tanpa analisis sintaksis yang tepat hasil terjemahan menjadi rancu. Oleh sebab itu dibutuhkan strategi pembelajaran yang efektif dan alat bantu yang mampu meminimalkan kesalahan semacam ini baik melalui pendekatan pedagogis maupun teknologi.

Seiring dengan kemajuan teknologi digital, bidang kecerdasan buatan (*Artificial Intelligence*) khususnya *Natural Language Processing* (NLP) mulai menawarkan solusi baru dalam pembelajaran bahasa dan penerjemahan[11]. Salah satu inovasi yang kini banyak dimanfaatkan adalah *Large Language Model* (LLM) seperti ChatGPT yang dikembangkan oleh OpenAI[12]. ChatGPT dirancang menggunakan arsitektur *transformer* yang mampu memahami konteks kalimat secara menyeluruh dengan memanfaatkan miliaran parameter dalam proses pembelajaran modelnya. Dengan kemampuan ini ChatGPT dapat menghasilkan terjemahan otomatis yang relatif baik dan lancar[13].

Penelitian-penelitian terbaru mulai menyoroti penggunaan ChatGPT dalam proses pembelajaran Bahasa Arab di Indonesia. Studi yang dilakukan oleh Saimin, Supriadi dan Al Farisi yang berjudul *Analisis Kesalahan Penerjemahan Teks Bahasa Indonesia ke dalam Bahasa Arab pada ChatGPT* menunjukkan bahwa terdapat beberapa kesalahan dan harus ditinjau kembali khususnya dalam tingkat morfologi dan sintaksis bahasa target. Sejalan dengan itu, penelitian yang dilakukan oleh Abdul Ruhmadi dan M. Zaka Al-Farisi menunjukkan bahwa terdapat kesalahan terjemahan bahasa Arab ke dalam bahasa Indonesia pada aspek morfologi dalam ChatGPT yang ditemukan di beberapa bagian[14]. Temuan serupa juga dikemukakan oleh Sinta Nailul Latifah dan Wulan Indah Fatimatul Djamilah menunjukkan bahwa ChatGPT mampu memberikan terjemahan yang cepat dan akurat dalam berbagai konteks akan tetapi tetap perlu di telaah ulang[15]. Berdasarkan telaah terhadap penelitian terdahulu dapat disimpulkan bahwa kajian yang ada masih berfokus pada analisis kesalahan morfologi dan sintaksis serta evaluasi kualitatif berbasis deskripsi linguistik saja.

Perguruan tinggi di Indonesia mulai mengintegrasikan ChatGPT sebagai alat pembelajaran interaktif dalam pembelajaran bahasa arab. Berdasarkan beberapa penelitian berskala institusi penggunaan ChatGPT sebagai pendukung pembelajaran bahasa Arab berada pada kisaran 68–72%. Persentase tersebut merujuk pada hasil survei persepsi dan minat mahasiswa terhadap penggunaan ChatGPT sebagai alat bantu pembelajaran bahasa Arab di tingkat perguruan tinggi. Mahasiswa memanfaatkan ChatGPT untuk menghasilkan ide, memeriksa hasil terjemahan, mendapatkan sinonim atau parafrase serta mengeksplorasi konteks kata[16]. Dosen juga memanfaatkan teknologi ini untuk memperkaya materi ajar dan memberikan umpan balik otomatis kepada mahasiswanya. Namun penggunaan AI dalam pembelajaran bahasa arab harus disertai dengan keterampilan berpikir kritis agar mahasiswa tidak sekadar menerima hasil terjemahan secara mentah tanpa mempertimbangkan konteks linguistik dan budaya yang melingkupinya. Ketergantungan berlebihan terhadap AI juga dapat menghambat kemampuan analisis bahasa dan kreativitas linguistik mahasiswa[17].

Meningkatnya pemanfaatan ChatGPT dalam pembelajaran dan penerjemahan Bahasa Arab menunjukkan bahwa teknologi kecerdasan buatan telah menjadi salah satu alat bantu yang semakin sering digunakan oleh mahasiswa dalam memahami teks berbahasa Arab. Kemudahan akses, kecepatan dalam menghasilkan terjemahan serta kemampuan ChatGPT dalam merespons berbagai jenis teks menjadikan teknologi ini banyak dimanfaatkan untuk

mendukung kegiatan akademik. Namun penggunaan yang semakin luas tersebut belum diiringi dengan pemahaman yang memadai mengenai kualitas hasil terjemahan yang dihasilkan. Dalam praktiknya sebagian pengguna cenderung menerima hasil terjemahan secara langsung tanpa melakukan verifikasi terhadap keakuratan makna, kesesuaian konteks, keberterimaan maupun keterbacaannya padahal Bahasa Arab memiliki karakteristik linguistik yang kompleks, mencakup aspek morfologi, sintaksis, semantik serta unsur budaya dan pragmatik yang dapat memengaruhi ketepatan hasil terjemahan. Kesalahan dalam menerjemahkan unsur-unsur tersebut berpotensi menimbulkan kesalahpahaman terhadap isi teks terutama dalam konteks akademik yang menuntut ketelitian dan ketepatan makna. Di sisi lain penelitian terdahulu mengenai penggunaan ChatGPT dalam penerjemahan Bahasa Arab masih lebih banyak berfokus pada identifikasi kesalahan linguistik dan analisis deskriptif terhadap hasil terjemahan. Penelitian yang mengombinasikan evaluasi otomatis menggunakan metrik BLEU dengan evaluasi manusia untuk memperoleh gambaran kualitas terjemahan secara lebih menyeluruh masih relatif terbatas. Akibatnya belum tersedia informasi yang komprehensif mengenai sejauh mana kualitas terjemahan ChatGPT dari Bahasa Arab ke Bahasa Indonesia serta bagaimana tingkat kesesuaian antara hasil penilaian mesin dan penilaian manusia. Oleh karena itu penelitian ini perlu dilakukan untuk memberikan evaluasi yang lebih objektif dan komprehensif terhadap kualitas terjemahan ChatGPT melalui pendekatan gabungan antara metrik BLEU dan evaluasi manusia sehingga hasilnya dapat menjadi dasar dalam pemanfaatan teknologi kecerdasan buatan secara lebih kritis, efektif dan bertanggung jawab dalam pembelajaran Bahasa Arab.

Penilaian terhadap kualitas terjemahan ChatGPT dapat dilakukan melalui dua pendekatan, yakni evaluasi otomatis dan evaluasi manusia. Evaluasi otomatis menggunakan metrik seperti BLEU, METEOR, BERTScore dan COMET. Metrik-metrik tersebut sering digunakan untuk menilai kesamaan antara terjemahan yang dihasilkan mesin dan teks referensi manusia secara kuantitatif [18]. Meskipun metrik-metrik ini menawarkan cara yang praktis dalam pengukuran namun penelitian terdahulu menunjukkan bahwa hasil skor evaluasi otomatis sering kali tidak sejalan dengan penilaian manusia terutama dalam aspek kealamian dan kesesuaian dengan konteks sosial dan budaya. Oleh karena itu evaluasi manusia masih dianggap menjadi metode yang paling akurat untuk menilai kualitas terjemahan dalam hal kesepadanan makna, gaya bahasa dan konteks budaya.

Dalam konteks pembelajaran Bahasa Arab di Indonesia penelitian terhadap kualitas terjemahan ChatGPT dianggap penting karena dapat memberikan perspektif yang luas mengenai sejauh mana teknologi ini dapat diandalkan sebagai alat bantu dalam pembelajaran bahasa arab. Penelitian semacam ini tidak hanya menilai kemampuan teknis ChatGPT dalam menghasilkan terjemahan yang akurat tetapi juga mengevaluasi sejauh mana model bahasa tersebut dalam menafsirkan nuansa pragmatik, idiomatik dan budaya yang terdapat dalam teks Arab. Guna memperoleh gambaran yang utuh diperlukan evaluasi yang mengintegrasikan antara metode evaluasi otomatis dan penilaian manusia. Melalui perpaduan kedua metode ini diharapkan dapat memperoleh pemahaman yang lebih menyeluruh yang menyeimbangkan aspek linguistik dan non-linguistik.

Meskipun penelitian sebelumnya telah mengkaji kinerja ChatGPT dalam menerjemahkan teks Bahasa Arab sebagian besar penelitian tersebut masih terbatas pada analisis kesalahan linguistik tanpa mengaitkannya dengan penilaian berbasis metrik otomatis dan penilaian manusia secara bersamaan. Oleh karena itu belum ada penelitian yang secara komprehensif menilai kualitas terjemahan ChatGPT dari Bahasa Arab ke Bahasa Indonesia menggunakan pendekatan gabungan tersebut. Penelitian ini memiliki kebaruan dalam hal pendekatan metodologisnya, yakni mengombinasikan evaluasi otomatis BLEU dengan evaluasi manusia untuk mendapatkan gambaran yang lebih komprehensif mengenai kualitas terjemahan yang belum dikaji secara sistematis dalam penelitian sebelumnya. Diharapkan kebaruan penelitian ini dapat berkontribusi secara eksperimental dalam pengembangan model pembelajaran bahasa Arab berbasis kecerdasan buatan di Indonesia. Berdasarkan uraian tersebut maka fokus penelitian ini adalah untuk mengetahui kualitas hasil terjemahan ChatGPT dari teks Bahasa Arab ke Bahasa Indonesia dengan menggunakan pendekatan gabungan antara metrik bleu dan evaluasi manusia sekaligus menganalisis tingkat kesesuaian kedua metode tersebut dalam menilai kualitas terjemahan ChatGPT sebagai alat terjemahan bahasa Arab.

II. METODE

Penelitian ini menggunakan metode kuantitatif yaitu metode yang bertujuan mendeskripsikan fenomena secara objektif melalui pengukuran numerik tanpa melakukan pengujian hubungan sebab-akibat[19]. Pemilihan metode ini didasarkan pada tujuan penelitian yang ingin menggambarkan kualitas hasil terjemahan ChatGPT dari teks Bahasa Arab ke Bahasa Indonesia berdasarkan skor penilaian evaluator dan nilai metrik otomatis sekaligus mengetahui sejauh mana kesesuaian kedua metode dalam menilai kualitas terjemahan ChatGPT sebagai alat bantu pembelajaran Bahasa Arab. Pendekatan kuantitatif dianggap paling sesuai karena peneliti tidak menganalisis hubungan antar variabel tetapi hanya menyajikan hasil pengukuran kualitas terjemahan secara statistik deskriptif sebagaimana dijelaskan oleh

Sugiyono bahwa penelitian deskriptif bertujuan untuk menggambarkan keadaan objek sesuai apa adanya berdasarkan data kuantitatif[20].

Objek penelitian ini mencakup teks Bahasa Arab yang digunakan dalam tugas mata kuliah Tarjamah Arab-Indonesia digital yang dipilih menggunakan teknik purposive sampling dengan kriteria pemilihan meliputi teks berbahasa Arab fusha, memuat struktur sintaksis yang beragam, mengandung unsur leksikal dan semantik yang menuntut pemahaman konteks, memiliki tingkat kompleksitas morfologis yang cukup untuk dianalisis kualitas terjemahannya dan teks berasal dari sumber berita Arab autentik. Subjek penelitian terdiri dari 30 mahasiswa program studi Pendidikan Bahasa Arab yang sudah menempuh mata kuliah tarjamah Arab-Indonesia Digital di Universitas Muhammadiyah Sidoarjo dan telah memahami dasar-dasar teori penerjemahan sebagai evaluator manusia.

Sumber data penelitian terdiri dari data primer dan data sekunder. Data primer meliputi hasil terjemahan ChatGPT, skor penilaian evaluator menggunakan skala Likert 1–5 terhadap tiga aspek utama kualitas terjemahan yakni Keakuratan, Keberterimaan dan Keterbacaan sebagaimana yang dipaparkan oleh Prof. Nababan[21] serta nilai metrik otomatis BLEU (Bilingual Evaluation Understudy). BLEU merupakan algoritma evaluasi terjemahan mesin yang membandingkan teks terjemahan kandidat dengan terjemahan referensi yang dibuat oleh manusia. Prinsip dasar BLEU adalah bahwa semakin tinggi tingkat kesamaan antara kedua teks tersebut maka semakin baik kualitas terjemahannya. Skor BLEU berada pada rentang 0 hingga 1 dengan skor mendekati 1 menunjukkan tingkat kemiripan dan akurasi yang tinggi[22]. Data sekunder dikumpulkan dari berbagai sumber seperti buku, artikel jurnal ilmiah serta penelitian terdahulu yang membahas linguistik arab, teori penerjemahan dan evaluasi penerjemahan otomatis.

Teknik pengumpulan data dalam penelitian ini dilakukan melalui beberapa tahapan secara sistematis. Tahap pertama adalah menentukan sampel penelitian berupa sembilan kalimat berbahasa Arab yang dipilih menggunakan teknik purposive sampling berdasarkan kriteria yang telah ditetapkan. Tahap kedua adalah melakukan penerjemahan teks Bahasa Arab menggunakan ChatGPT. Seluruh teks dimasukkan ke dalam ChatGPT dengan instruksi yang sama untuk menjaga konsistensi hasil terjemahan. Hasil terjemahan kemudian didokumentasikan dan disimpan sebagai data penelitian. Tahap ketiga adalah menyusun terjemahan rujukan (reference translation) yang digunakan sebagai pembanding dalam evaluasi otomatis. Terjemahan rujukan dibuat berdasarkan kaidah penerjemahan Bahasa Arab–Indonesia yang baik dan benar dengan mempertimbangkan kesepadanan makna, struktur kalimat dan konteks bahasa sasaran. Tahap keempat adalah pengumpulan data untuk evaluasi otomatis menggunakan metrik BLEU. Pada tahap ini hasil terjemahan ChatGPT dibandingkan dengan terjemahan rujukan menggunakan bahasa pemrograman Python. Proses penghitungan dilakukan dengan membandingkan kesamaan n-gram antara kedua teks sehingga diperoleh skor BLEU untuk setiap kalimat dan skor rata-rata keseluruhan. Tahap kelima adalah penyusunan instrumen evaluasi manusia berdasarkan tiga aspek kualitas terjemahan yang dikemukakan oleh Nababan yaitu keakuratan, keberterimaan dan keterbacaan. Instrumen disusun dalam bentuk angket menggunakan skala Likert 1–5 yang terdiri atas 11 butir pernyataan. Tahap keenam adalah penyebaran instrumen kepada 30 mahasiswa Program Studi Pendidikan Bahasa Arab Universitas Muhammadiyah Sidoarjo yang telah menempuh mata kuliah Tarjamah Arab–Indonesia Digital. Para mahasiswa diminta membaca hasil terjemahan ChatGPT kemudian memberikan penilaian terhadap setiap pernyataan yang terdapat dalam angket. Tahap ketujuh adalah pengumpulan dan dokumentasi seluruh data penelitian yang meliputi teks sumber berbahasa Arab, hasil terjemahan ChatGPT, terjemahan rujukan, skor BLEU serta hasil penilaian responden. Seluruh data kemudian diolah menggunakan SPSS untuk keperluan analisis statistik deskriptif.

Data penelitian dianalisis menggunakan teknik statistik deskriptif dan analisis komparatif. Analisis untuk menjawab rumusan masalah pertama mengenai kualitas terjemahan ChatGPT berdasarkan metrik BLEU dilakukan dengan cara menghitung skor BLEU pada setiap kalimat menggunakan bahasa pemrograman Python. Selanjutnya dihitung nilai rata-rata, nilai minimum, nilai maksimum dan rentang skor untuk menggambarkan tingkat kesesuaian antara hasil terjemahan ChatGPT dengan terjemahan rujukan. Analisis untuk menjawab rumusan masalah kedua mengenai kualitas terjemahan berdasarkan evaluasi manusia dilakukan dengan menghitung nilai rata-rata (mean), nilai minimum, nilai maksimu dan standar deviasi menggunakan program SPSS. Hasil perhitungan kemudian diinterpretasikan berdasarkan kategori penilaian Sugiyono yaitu 1,00–1,80 (sangat rendah), 1,81–2,60 (rendah), 2,61–3,40 (sedang), 3,41–4,20 (tinggi) dan 4,21–5,00 (sangat tinggi). Selanjutnya analisis kesesuaian antara metrik BLEU dan evaluasi manusia dilakukan menggunakan pendekatan deskriptif-komparatif dengan membandingkan kecenderungan hasil kedua metode. Analisis ini bertujuan untuk mengetahui sejauh mana hasil evaluasi otomatis sejalan dengan hasil evaluasi manusia dalam menilai kualitas terjemahan ChatGPT.

Hasil analisis disajikan dalam bentuk tabel dan deskripsi naratif guna memberikan gambaran objektif mengenai kualitas terjemahan ChatGPT sedangkan Validitas instrumen dilakukan melalui validitas konten dengan melibatkan dosen ahli yang mengevaluasi sejauh mana indikator-indikator tersebut mewakili pengukuran kualitas terjemahan. Penggunaan prosedur ini memastikan bahwa data yang diperoleh memiliki tingkat akurasi dan konsistensi yang memadai. Dengan demikian metode kuantitatif deskriptif ini mampu memberikan gambaran yang sistematis, objektif dan komprehensif mengenai kualitas terjemahan ChatGPT serta potensi pemanfaatannya sebagai alat bantu pembelajaran Bahasa Arab.

III. HASIL DAN PEMBAHASAN

1. Kualitas Terjemahan ChatGPT Berdasarkan Metrik BLEU

Sebagai bentuk evaluasi otomatis terhadap kualitas terjemahan penelitian ini memanfaatkan metrik BLEU yang banyak digunakan dalam penelitian penerjemahan mesin. Penggunaan metrik ini bertujuan untuk mengukur tingkat kesamaan antara hasil terjemahan ChatGPT dan terjemahan rujukan yang disusun oleh manusia. Hasil pengukuran tersebut kemudian digunakan sebagai dasar untuk menilai kualitas terjemahan secara kuantitatif**.** Penghitungan skor BLEU dalam penelitian ini dilakukan melalui beberapa tahapan yaitu data berupa teks Bahasa Arab yang telah diterjemahkan menggunakan ChatGPT dibandingkan dengan terjemahan rujukan (hasil terjemahan manusia) kemudian teks dinormalisasi yaitu dengan menjadikan teks seragam seperti mengubah teks menjadi huruf kecil dan menghapus spasi kosong yang tidak diperlukan. Setelah itu teks dibagi menjadi beberapa unit kalimat agar penghitungan skor BLEU dapat dilakukan secara lebih mudah dan akurat[23]. Teks yang telah dibagi menjadi beberapa unit kalimat kemudian dihitung secara otomatis menggunakan bahasa pemrograman Python 3.14 melalui Command Prompt. Hasil penghitungan tersebut menghasilkan skor BLEU pada setiap kalimat serta nilai rata-rata keseluruhan yang digunakan untuk mengetahui tingkat kesesuaian terjemahan. Sebagai ilustrasi proses evaluasi kualitas terjemahan berikut disajikan salah satu contoh data yang digunakan dalam penelitian.

Teks Bahasa Arab [24]

كما استشهد سابق اليوم محمد خالد عابد برصاص إسرائيلي في الرأس بمنطقة الزرقاء في حي التفاح وفقاً للمصدر الطبي

Hasil Terjemahan ChatGPT

Muhammad Khalid Abid gugur hari ini akibat tembakan Israel di kepala di wilayah Az-Zarqa di lingkungan At-Tuffah menurut sumber medis.

Terjemahan Rujukan

Menurut sumber medis, Muhammad Khalid Abid tewas akibat tembakan pasukan Israel di kepala di kawasan Az-Zarqa, permukiman At-Tuffah, hari ini. Pada contoh di atas terdapat perbedaan struktur sintaksis dan pilihan leksikal antara hasil terjemahan ChatGPT dan terjemahan rujukan. Perbedaan tersebut menyebabkan tingkat kesamaan n-gram menjadi rendah sehingga menghasilkan skor BLEU yang relatif kecil meskipun makna umum masih dapat dipahami. Adapun hasil perhitungan skor BLEU secara rinci pada setiap kalimat dapat dilihat pada tabel berikut :

Tabel 1. Hasil analisis metrik BLEU

No	Kalimat	Skor Bleu
1.	Kalimat 1	0.3583
2.	Kalimat 2	0.4010
3.	Kalimat 3	0.4980
4.	Kalimat 4	0.4204
5.	Kalimat 5	0.0764
6.	Kalimat 6	0.4896
7.	Kalimat 7	0.4422
8.	Kalimat 8	0.4249
9.	Kalimat 9	0.6004
Rata-Rata Bleu		**0.4123**

Tabel 1 menyajikan hasil perhitungan nilai BLEU yang menunjukkan adanya variasi skor yang cukup signifikan. Nilai minimum sebesar 0,0764 yang terdapat pada kalimat nomor 5 sedangkan nilai maksimum sebesar 0,6004 terdapat pada kalimat nomor 9 dimana secara konseptual skor BLEU berada pada rentang 0–1 dimana nilai yang lebih tinggi menunjukkan tingkat kesesuaian n-gram yang lebih besar antara terjemahan mesin dan terjemahan rujukan[25]. Namun penting ditegaskan bahwa hingga saat ini tidak terdapat standar kategorisasi baku yang secara universal mengklasifikasikan skor BLEU ke dalam kategori seperti rendah, sedang atau tinggi [26] oleh karena itu interpretasi skor dalam penelitian ini tidak didasarkan pada kategorisasi absolut melainkan pada pendekatan komparatif dan kontekstual. Dalam pendekatan komparatif terdapat selisih yang cukup besar antara nilai minimum dan maksimum yang menghasilkan rentang skor sebesar 0,5240 yang mengindikasikan adanya kesenjangan kualitas terjemahan antar kalimat. Skor yang sangat rendah pada kalimat tertentu menunjukkan rendahnya tingkat kesesuaian n-gram yang dapat diinterpretasikan sebagai adanya perbedaan signifikan antara hasil terjemahan ChatGPT dan

terjemahan rujukan baik dari segi struktur sintaksis maupun leksikal. Sebaliknya skor yang lebih tinggi menunjukkan bahwa dalam kondisi tertentu ChatGPT mampu menghasilkan terjemahan yang lebih mendekati rujukan[18] Temuan ini sejalan dengan teori penerjemahan yang dikemukakan oleh Nida bahwa kesepadanan terjemahan tidak hanya bersifat formal (formal equivalence) tetapi juga harus memperhatikan kesepadanan dinamis (dynamic equivalence) yaitu sejauh mana pesan dalam bahasa sumber dapat diterima secara alami dalam bahasa sasaran[27]. Dengan demikian rendahnya skor BLEU tidak selalu menunjukkan kegagalan makna secara keseluruhan melainkan dapat disebabkan oleh perbedaan struktur dan pilihan ekspresi yang tidak tertangkap oleh perhitungan n-gram.

Sebagai contoh pada kalimat nomor lima dengan skor terendah yang berbunyi “Kama istashhada sabiqal-yaum, Muhammad Khalid Abid birasas Israili fi ar-ra's bimantiqah az-Zarqa fi hayy at-Tuffah wifqa al-masdar at-tibbi” pada kalimat tersebut terlihat adanya ketidaksesuaian dari aspek sintaksis dan leksikal. Secara sintaksis frasa “kama istashhada sabiqal-yaum” merupakan struktur jumlah fi’liyah yang diawali oleh keterangan peristiwa dan waktu (dzarf zaman) jika diterjemahkan secara harfiah menjadi “sebagaimana gugur sebelumnya hari ini” terjemahan tersebut terasa kaku dan tidak alami. Hal ini disebabkan karena adanya perbedaan pola dasar antara bahasa Arab yang cenderung fleksibel dalam penempatan unsur keterangan dibandingkan bahasa Indonesia yang lebih mengutamakan pola Subjek Predikat Objek (S-P-O) serta kejelasan alur informasi. Selain itu pada frasa “wifqa al-masdar at-tibbi” jika diterjemahkan “menurut sumber medis” secara leksikal sudah tepat namun dalam praktik penulisan bahasa Indonesia frasa tersebut lebih lazim ditempatkan di awal kalimat untuk meningkatkan keterbacaan. Hal ini sejalan dengan pendapat Newmark yang menekankan pentingnya penyesuaian struktur dalam bahasa sasaran agar menghasilkan terjemahan yang komunikatif (communicative translation)[28]. Oleh karena itu terjemahan yang lebih alami dapat direkonstruksi menjadi “Menurut sumber medis, Muhammad Khalid Abid tewas akibat tembakan di kepala oleh pasukan Israel di kawasan Az-Zarqa permukiman At-Tuffah, hari ini.”

Sementara itu dalam pendekatan kontekstual variasi skor tersebut menunjukkan bahwa kualitas terjemahan tidak bersifat konsisten dan sangat dipengaruhi oleh karakteristik linguistik teks sumber baik kompleksitas struktur kalimat bahasa Arab, ambiguitas makna leksikal maupun perbedaan sistem gramatikal antara bahasa sumber dan bahasa sasaran yang berpotensi mempengaruhi tingkat kesesuaian hasil terjemahan. Rentang skor yang lebar ini memperkuat indikasi bahwa performa sistem masih fluktuatif dalam menangani variasi struktur dan makna. Temuan ini sejalan dengan pandangan Larson yang menegaskan bahwa penerjemahan tidak hanya berorientasi pada pengalihan bentuk bahasa tetapi juga pada penyampaian makna yang kontekstual dan alami dalam bahasa sasaran[29]. Perlu dipahami bahwa BLEU sebagai metrik evaluasi berbasis n-gram memiliki keterbatasan karena hanya mengukur kesamaan pada permukaan dan belum sepenuhnya mampu mempertimbangkan makna secara kontekstual[30] oleh karena itu skor BLEU yang diperoleh dalam penelitian ini tidak dapat dijadikan satu-satunya indikator kualitas terjemahan melainkan perlu dilengkapi dengan evaluasi manusia untuk memperoleh gambaran yang lebih komprehensif mengenai keakuratan, keberterimaan dan keterbacaan hasil terjemahan.

2. Kualitas Terjemahan ChatGPT Berdasarkan Evaluasi Manusia

Selain evaluasi otomatis menggunakan metrik BLEU, penelitian ini juga melibatkan evaluasi manusia untuk memperoleh gambaran yang lebih komprehensif mengenai kualitas terjemahan ChatGPT. Evaluasi manusia dilakukan karena kualitas terjemahan tidak hanya ditentukan oleh kesamaan bentuk bahasa tetapi juga oleh aspek keakuratan makna, keberterimaan dan keterbacaan yang hanya dapat dinilai secara optimal oleh pembaca manusia. Adapun data hasil penyebaran kuesioner (Evaluasi Manusia) pada 30 Mahasiswa PBA Umsida yang terdiri atas 11 butir pernyataan mengenai keakuratan, keberterimaan dan keterbacaan terjemahan ChatGpt diolah dengan bantuan program SPSS untuk mengetahui nilai rata-rata (mean), nilai minimum, nilai maksimum serta standar deviasi dari setiap butir pernyataan yang kemudian dianalisis berdasarkan interpretasi Sugiyono dengan interval 1,00–1,80 (sangat rendah), 1,81–2,60 (rendah), 2,61–3,40 (sedang), 3,41–4,20 (tinggi) dan 4,21–5,00 (sangat tinggi) yang disajikan dalam bentuk tabel berikut ini :

Tabel 2. Hasil analisis statistik deskriptif

	N	Range	Minimum	Maximum	Mean	Std. Deviation
satu	30	3	2	5	3.50	.900
dua	30	3	2	5	4.20	.610
tiga	30	4	1	5	2.87	1.106
empat	30	3	2	5	3.63	.809
lima	30	3	2	5	3.50	.900
enam	30	4	1	5	3.13	1.167
tujuh	30	4	1	5	3.53	1.008
delapan	30	4	1	5	2.73	1.112

sembilan	30	3	2	5	4.23	.626
sepuluh	30	2	3	5	3.83	.699
sebelas	30	3	2	5	3.07	1.081
Valid N (listwise)	30					

Tabel 2 menunjukkan bahwa dari 30 responden yang memberikan jawaban terhadap instrumen yang telah disebar diperoleh nilai rata-rata pada aspek keakuratan yang terdapat pada item nomor 1-4 berkisar antara 2,87 hingga 4,20 dengan rata-rata keseluruhan sebesar 3,55 yang termasuk dalam kategori tinggi. Hal ini menunjukkan bahwa terjemahan yang dihasilkan oleh ChatGPT secara umum dinilai cukup akurat oleh responden. Sedangkan pada aspek keberterimaan yang terdapat pada item nomor 5-7 nilai rata-ratanya berkisar pada rentang 3,13 hingga 3,53 dengan rata-rata keseluruhan sebesar 3,39 yang termasuk pada kategori sedang. Hal ini mengindikasikan bahwa hasil terjemahan masih cukup dapat diterima meskipun terdapat beberapa bagian yang kurang sesuai dengan kaidah bahasa sasaran. Sementara itu pada aspek keterbacaan yang terdapat pada item nomor 8-11 nilai rata-ratanya berkisar antara 2,73 hingga 4,23 dengan rata-rata keseluruhan sebesar 3,47 yang termasuk dalam kategori tinggi. Hal ini menunjukkan bahwa teks hasil terjemahan relatif mudah dipahami oleh pembaca.

Responden juga memberikan catatan bahwa penggunaan ChatGPT sebagai alat penerjemahan dinilai cukup membantu dalam memahami teks khususnya dari Bahasa Arab ke Bahasa Indonesia. Mayoritas responden menyatakan bahwa hasil terjemahan telah memenuhi aspek keterbacaan dan cukup mudah dipahami meskipun pada beberapa bagian masih memerlukan pembacaan ulang untuk memperoleh pemahaman yang lebih jelas. Adapun dari segi kualitas, terjemahan yang dihasilkan cenderung telah menyampaikan makna secara umum atau akurat namun masih ditemukan beberapa ketidaktepatan dalam pemilihan diksi serta kurangnya kesesuaian konteks pada bagian tertentu. Selain itu dari aspek keberterimaan hasil terjemahan sering kali dinilai kurang alami karena penggunaan bahasa yang terlalu baku dan kaku sehingga terkesan sebagai hasil terjemahan mesin. Hal ini menunjukkan bahwa meskipun struktur kalimat secara gramatikal dapat diterima sedangkan tingkat kealamiahan bahasa masih perlu ditingkatkan. Responden juga menyoroti bahwa kualitas terjemahan dipengaruhi oleh kemampuan pengguna dalam memahami konteks bahasa sumber dan bahasa sasaran. Berdasarkan tanggapan responden diatas dapat disimpulkan bahwa ChatGPT memiliki potensi yang cukup baik sebagai alat bantu penerjemahan namun penggunaannya masih memerlukan evaluasi dan penyuntingan lanjutan oleh manusia untuk mencapai kualitas terjemahan yang optimal terutama dalam aspek keberterimaan dan ketepatan konteks.

3. Kesesuaian Metrik BLEU dan Evaluasi Manusia

Setelah dilakukan analisis kualitas terjemahan melalui metrik BLEU dan evaluasi manusia tahap berikutnya adalah mengkaji bagaimana hubungan dan kesesuaian hasil yang diperoleh dari kedua metode tersebut. Analisis ini dilakukan untuk memberikan gambaran yang lebih utuh mengenai kualitas terjemahan ChatGPT serta mengetahui sejauh mana penilaian yang dihasilkan oleh evaluasi otomatis sejalan dengan penilaian yang diberikan oleh manusia. Tingkat kesesuaian antara Metrik BLEU dan Evaluasi manusia dalam penelitian ini tidak dianalisis menggunakan uji korelasi pearson disebabkan oleh perbedaan unit analisis pada kedua metode dimana skor metrik BLEU dihitung berdasarkan unit kalimat (9 kalimat) sedangkan evaluasi manusia diperoleh dari instrumen skala Likert yang terdiri atas 11 butir pernyataan yang mengukur aspek keakuratan, keberterimaan dan keterbacaan. Perbedaan struktur data tersebut menyebabkan kedua variabel tidak dapat dibandingkan secara langsung dalam bentuk pasangan data yang sepadan sehingga uji korelasi tidak memenuhi asumsi metodologis. Oleh karena itu analisis kesesuaian antara metrik BLEU dan Evaluasi manusia dalam penelitian ini menggunakan pendekatan deskriptif-komparatif dengan membandingkan kecenderungan hasil dari masing-masing metode. Hasil analisis menunjukkan bahwa skor BLEU yang diperoleh cenderung bersifat fluktuatif dengan rentang yang cukup lebar sementara hasil evaluasi manusia menunjukkan kecenderungan penilaian yang relatif stabil pada kategori sedang hingga tinggi. Perbedaan ini mengindikasikan bahwa kedua metode tidak selalu memberikan penilaian yang selaras terhadap kualitas terjemahan. Secara lebih spesifik terdapat kondisi di mana skor BLEU yang rendah tidak selalu diikuti oleh penilaian manusia yang rendah. Hal ini menunjukkan bahwa meskipun secara leksikal terjemahan memiliki tingkat kesamaan yang rendah terhadap teks rujukan akan tetapi secara makna masih dapat dipahami dengan baik oleh pembaca. Sebaliknya skor BLEU yang tinggi tidak selalu menjamin bahwa terjemahan tersebut memiliki tingkat keberterimaan yang tinggi terutama dalam aspek kealamiahan bahasa. Temuan ini sejalan dengan Papineni yang menyatakan bahwa BLEU hanya mengukur kesamaan n-gram dan tidak mempertimbangkan makna secara kontekstual[18] oleh karena itu evaluasi manusia tetap diperlukan untuk menilai aspek semantik dan pragmatik dalam terjemahan.

Dengan demikian dapat disimpulkan bahwa tingkat kesesuaian antara metrik BLEU dan evaluasi manusia dalam penelitian ini bersifat tidak langsung dan parsial sehingga kedua metode lebih tepat digunakan secara

komplementer daripada komparatif statistik. Pendekatan ini menunjukkan bahwa evaluasi kualitas terjemahan yang komprehensif memerlukan kombinasi antara metode otomatis dan penilaian manusia.

IV. SIMPULAN

Penelitian ini menunjukkan bahwa kualitas terjemahan ChatGPT dari Bahasa Arab ke Bahasa Indonesia berada pada tingkat yang cukup baik namun bersifat fluktuatif. Hasil Evaluasi otomatis menggunkan metrik BLEU Memperoleh skor rata-rata 0.4123 yang mencerminkan tingkat moderat antara terjemahan mesin dan terjemahan rujukan dengan rentang variasi yang cukup lebar antar kalimat. Variasi tersebut mengindikasikan bahwa performa ChatGpt sangat dipengaruhi oleh kompleksitas struktur linguistik khususnya pada aspek sintaksis, pemilihan leksikal serta perbedaan sistem gramatikal antara bahasa sumber dan bahasa sasaran. Adapun hasil evaluasi manusia memperlihatkan bahwa kualitas terjemahan berada pada kategori tinggi pada aspek keakuratan dan keterbacaan sedangkan pada aspek keberterimaan berada dalam kategori sedang. Temuan ini menunjukkan bahwa terjemahan yang dihasilkan telah mampu menyampaikan makna utama dengan tingkat pemahaman yang relatif baik namun masih ditemukan kecenderungan penggunaan struktur kalimat yang kurang alami dan pemilihan diksi yang sepenuhnya sesuai dengan kaidah Bahasa Indonesia.

Penelitian ini juga menemukan bahwa hubungan antara skor BLEU dan penilaian manusia tidak bersifat linier yang menegaskan keterbatasan BLEU dalam menangkap dimensi semantik dan pragmatik. Oleh karena itu evaluasi kualitas terjemahan perlu dilakukan secara komplementer dengan memadukan metode otomatis dan penilaian manusia. Implikasi penelitian ini menunjukkan bahwa ChatGPT dapat diintegrasikan dalam pembelajaran Bahasa Arab sebagai media pembelajaran berbasis teknologi yang bersifat adaptif dan interaktif khususnya dalam mata kuliah tarjamah. Pemanfaatannya tidak hanya sebagai alat penerjemah saja tetapi juga sebagai sarana untuk mengembangkan keterampilan berpikir kritis melalui aktivitas analisis, evaluasi dan revisi terjemahan.

V. SARAN

Berdasarkan temuan penelitian ini terdapat beberapa rekomendasi yang dapat diajukan sebagai upaya dalam pengembangan penelitian selanjutnya yaitu : Bagi peneliti selanjutnya disarankan untuk menggunakan unit analisis yang sepadan antara evaluasi otomatis dan evaluasi manusia sehingga memungkinkan dilakukan uji statistik inferensial seperti korelasi atau regresi. Selain itu penggunaan metrik evaluasi lain seperti METEOR, BERTScore atau COMET juga dapat dipertimbangkan untuk memperoleh hasil yang lebih variatif dan mendalam. Penelitian selanjutnya juga dapat memperluas cakupan data dengan melibatkan jenis teks yang lebih beragam seperti teks sastra, teks turats maupun teks media modern guna menguji konsistensi performa ChatGPT dalam berbagai konteks linguistik. Dengan mempertimbangkan hal tersebut diharapkan pemanfaatan teknologi kecerdasan buatan dalam pembelajaran Bahasa Arab dapat dilakukan secara lebih optimal, kritis dan proporsional.

UCAPAN TERIMA KASIH

Alhamdulillahi rabbil ‘alamin, segala puji dan syukur penulis panjatkan ke hadirat Allah SWT atas segala rahmat, karunia serta pertolongan-Nya sehingga penelitian ini dapat diselesaikan dengan baik. Penulis menyampaikan terima kasih kepada seluruh pihak yang telah memberikan dukungan dalam penyelesaian penelitian ini. Ucapan terima kasih secara khusus disampaikan kepada dosen pembimbing atas arahan, bimbingan serta masukan yang konstruktif selama proses penelitian. Penulis juga mengapresiasi partisipasi mahasiswa Program Studi Pendidikan Bahasa Arab Universitas Muhammadiyah Sidoarjo yang telah berkontribusi sebagai responden dalam penelitian ini. Semoga penelitian ini dapat memberikan manfaat dan kontribusi bagi pengembangan pembelajaran Bahasa Arab khususnya dalam pemanfaatan teknologi kecerdasan buatan.

REFERENSI

[1] R. Najwa, F. Azkiya, and T. Amelia, “Faktor-Faktor Kesulitan Mahasiswa : Pengaruh Gramatikal Bahasa Arab Terhadap Pembuatan Teks Bahasa Arab Factors of Student Difficulties : The Influence of Arabic

Grammar on Arabic Text Creation," pp. 11590–11597, 2025.

[2] K. Peneliti, *Bahasa arab di Indonesia*. 1436. [Online]. Available: https://ketabpedia.com/تحميل/اللغة-العربية-في-إندونيسيا/

[3] "Pelajaran Bahasa Arab untuk Non-Penutur Asli - Bagian Pertama.pdf." [Online]. Available: https://archive.org/details/20191011_20191011_0820/دروس اللغة العربية لغير الناطقين بها - الجزء الأول/

[4] L. Pertiwi, F. Zahara, J. K. Siagian, and H. A. Nasution, "Pengaruh Status Sosial terhadap Pilihan Dialek Bahasa Arab dalam Komunikasi Sehari-hari," *Al-Lughoh J. Bhs. Arab dan Sastra Arab*, vol. 1, no. 1, pp. 7–15, 2025.

[5] Ainur Rofiq Sofa and Ayun Febrianti, "Dialektologi Bahasa Arab: Analisis Perbedaan Linguistik Berdasarkan Kajian Pustaka," *Lencana J. Inov. Ilmu Pendidik.*, vol. 3, no. 2, pp. 76–87, 2025, doi: 10.55606/lencana.v3i2.5018.

[6] A. N. Fradana, *Buku Ajar Morfologi Bahasa*. 2018. doi: 10.21070/2018/978-602-5914-31-7.

[7] F. Rizka and F. M. Ammar, "Analisis Faktor Kesulitan Membaca Teks Berbahasa Arab Kelas VIII," *JIIP - J. Ilm. Ilmu Pendidik.*, vol. 7, no. 4, pp. 3660–3666, 2024, doi: 10.54371/jiip.v7i4.4295.

[8] M. Z. Haq, "Arabic Idiomatic Translation Problems To Indonesian Problematika Penerjemahan Idiomatik Arab Ke Indonesia," *MUHIBBUL Arab. J. Pendidik. Bhs. Arab*, vol. 2, no. 1, pp. 15–30, 2022.

[9] M. I. Salim and L. M. A. KH. M. Abd. Rouf, *Syarah Diwan Imam Asy-Syafi'i*. Divapress, 2019. [Online]. Available: https://books.google.co.id/books?id=eJ27DwAAQBAJ

[10] A. Marhamah, Imas Marliana, Halimah Ibrahim, and Sahkholid Nasution, "Analisis Kesalahan Linguistik Dalam Penerjemahan Teks Bahasa Arab Pada Google Translate," *AL-MUADDIB J. Kaji. Ilmu Kependidikan*, vol. 7, no. 1, pp. 122–136, 2025, doi: 10.46773/muaddib.v7i1.1513.

[11] S. Rifky *et al.*, *Artificial Intelligence : Teori dan Penerapan AI di Berbagai Bidang*. PT. Sonpedia Publishing Indonesia, 2024. [Online]. Available: https://books.google.co.id/books?id=r8YMEQAAQBAJ

[12] Ibrahim Ahmad Assegaf *et al.*, "Pengembangan Chatbot Konsultasi Kesehatan Mental Kesehatan Mental Berbasis Open Ai Model Gpt-3.5 Turbo Menggunakan Media Whatsapp," *J. Inform. Teknol. dan Sains*, vol. 6, no. 4, pp. 785–793, 2024, doi: 10.51401/jinteks.v6i4.4749.

[13] L. Judijanto *et al.*, *Optimalisasi ChatGPT: Panduan dan Penerapan untuk Belajar, Mengajar, dan Membuat Konten Tanpa Batas*. PT. Green Pustaka Indonesia, 2025. [Online]. Available: https://books.google.co.id/books?id=mhBlEQAAQBAJ

[14] A. Ruhmadi and M. Z. Al Farisi, "Analisis Kesalahan Morfologi Penerjemahan Arab–Indonesia pada ChatGPT," *Aphorisme J. Arab. Lang. Lit. Educ.*, vol. 4, no. 1, pp. 55–75, 2023, doi: 10.37680/aphorisme.v4i1.3148.

[15] S. N. Latifah and W. I. F. Djamilah, "Penggunaan chatgpt dalam penerjemahan bahasa arab," *Al Dirosah J. Pendidik. Agama Islam*, vol. 01, no. 2, pp. 1–13, 2024, [Online]. Available: http://ojs.iai-darussalam.ac.id/index.php/pai

[16] E. M. Z, "Minat Mahasiswa Menggunakan ChatGPT dalam Pembelajaran Bahasa," vol. 5, no. 1, pp. 155–168, 2025.

[17] A. S. M. Amadi and K. Hikmah, "Persepsi Mahasiswa Tentang Pemanfaatan Teknologi AI dalam

Pembelajaran Bahasa Arab di Perguruan Tinggi Islam Indonesia," *J. Educ. Res.*, vol. 6, no. 2, pp. 291–301, 2025, doi: 10.37985/jer.v6i2.2343.

[18] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, vol. 2002-July, no. July, pp. 311–318, 2002.

[19] M. Waruwu, S. N. Pu`at, P. R. Utami, E. Yanti, and M. Rusydiana, "Metode Penelitian Kuantitatif: Konsep, Jenis, Tahapan dan Kelebihan," *J. Ilm. Profesi Pendidik.*, vol. 10, no. 1, pp. 917–932, 2025, doi: 10.29303/jipp.v10i1.3057.

[20] S. P. M. A. D. A. B. M. P. Dr. Ishak Bagea, *Metode Penelitian Kuantitatif Kualitatif Campuran R \& D*. CV. AZKA PUSTAKA, 2025. [Online]. Available: https://books.google.co.id/books?id=KI5yEQAAQBAJ

[21] M. T. Kholida, "Penilaian Kualitas Keterbacaan Terjemahan Google Translate pada Surah Al- Alaq," vol. 1, pp. 98–104, 2025.

[22] H. A. Karim, "Journal of Digital Business and Information Technology," pp. 50–60, doi: 10.23971/jobit.v1i2.297.

[23] I. Yadi, "Implentasi Arsitektur Transfomer untuk penerjemahan Bahasa Lampung - Indonesia," 2024.

[24] "Kama istusyhida sabiqa al-yaum Muḥammad Khalid 'Abid bi-rasasin Isra'iliyyin fī ar-ra's bi-mantiqati az-Zarqa'' fī hayyi at-Tuffaḥ wafqan lil-masdari at-tibbi.,'" Al jazeera. [Online]. Available: https://www.aljazeera.net/news/2026/1/26/3-شهداء-بنيران-الاحتلال-في-مناطق-بقطاع

[25] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "B LEU : a Method for Automatic Evaluation of Machine Translation," no. July, pp. 311–318, 2002.

[26] M. Post, "A Call for Clarity in Reporting BLEU Scores ACL Anthology ID W18-6319 / revision 9 / 18 Sep 2025," vol. 1, pp. 186–191, 2018.

[27] A. Muam, C. D. Nugraha, and U. G. M. Press, *Pengantar Penerjemahan*. Gadjah Mada University Press, 2021. [Online]. Available: https://books.google.co.id/books?id=U3cWEAAAQBAJ

[28] Z. S. S. Hariyanto, *Translation: Bahasan Teori \& Penuntun Praktis Menerjemahkan*. Media Nusa Creative (MNC Publishing). [Online]. Available: https://books.google.co.id/books?id=LYZOEAAAQBAJ

[29] Y. R. Uyuni, *Menerjemahkan Makna Bukan Kata: Teori dan Evaluasi Penerjemahan Arab-Indonesia*. Penerbit A-Empat, 2023. [Online]. Available: https://books.google.co.id/books?id=0kPAEAAAQBAJ

[30] M. Julkarnain, E. Mardinata, and R. Susilowati, "Penerapan Transformer-Based Neural Machine Translation untuk Bahasa Bima," pp. 793–802, 2025, doi: 10.33364/algoritma/v.22-2.2949.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.