Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Systematic Review and Meta-Analysis of 2024–2025 Studies on AI Arabic Translation, Linguistics and Pedagogy

Reima Al-Jarf
Full Professor of English and Translation Studies, Riyadh, Saudi Arabia
Corresponding Author: Reima Al-Jarf**, E-mail**: reima.al.jarf@gmail.com

| ABSTRACT

This study aims to conduct a systematic review (SR) and meta-analysis (MA) of twenty articles by the author published between 2024–2025 on the use of AI models such as Microsoft Copilot (MC), DeepSeek (DS), and Google Translate (GT) in translation, linguistics, and education. It aims to answer the following question: What do the author's 2024–2025 AI studies collectively reveal about AI performance, limitations, and sociolinguistic implications? After identifying the corpus, applying eligibility (inclusion and exclusion) criteria, describing corpus characteristics, information sources, study design, data extraction, quality assessment, meta-analysis procedures, and data synthesis, and presenting the PRISMA flow, the articles were classified into seven thematic clusters: AI translation of technical terms and metaphorical expressions; AI and phonological processing; AI's ability to recognize and decode Arabic, Japanese, and Chinese calligraphic text images; AI interaction with real-world discourse, political language, and encrypted communication; AI and student translators; linguistic competence of four AI models and the different errors they produce; and human attitudes toward AI and academic practices. The SR and MA results revealed varied AI translation performance across domains, with technical terms, metaphorical expressions, and culturally embedded terminology presenting distinct challenges. Subclusters highlight differences in domain-specific terminology, structural patterns, folk expressions, and diachronic shifts. AI struggled with emphatic negation and full-text academic discourse but showed moderate success in chemical translation compared to human translators. Phonological interpretation and encrypted Arabic posed additional complexity, while calligraphic decoding and linguistic reasoning showed partial recognition. Temporal patterns also emerged: studies from early 2025 showed lower accuracy and higher hallucination rates, while those from late 2025 reflected improved performance, likely due to incremental model updates and domain accessibility. The MA used proportion-based effect sizes and a mixed-methods synthesis, combining quantitative accuracy measures with qualitative discourse analysis. Finally, human attitudes toward AI-generated academic work ranged from cautious acceptance to critical rejection, shaped by perceived quality, ethical concerns, and disciplinary norms. Together, the findings offer a nuanced understanding of AI's evolving linguistic behavior and its implications for translation pedagogy, academic integrity, and cross-cultural communication.

| KEYWORDS

Meta-analysis, systematic review, AI translation studies, 2024-2025 studies, Al-Jarf's AI research, linguistic performance, AI error analysis, AI calligraphic text recognition, AI phonological processing, translation pedagogy

| ARTICLE INFORMATION

ACCEPTED: 01 January 2026 **PUBLISHED:** 03 January 2026 **DOI:** 10.32996/jcsts.2026.5.1.2

1. Introduction

A systematic review (SR)[1] is a broad, rigorous process of finding, appraising, and summarizing all relevant research works on a topic. It is a comprehensive and transparent search and evaluation of studies using strict methods to answer a specific question by finding, evaluating, and synthesizing all available evidence, with predefined criteria. It involves clear research questions (PICO), extensive literature searches, quality assessment, and synthesis of findings. The outcome is a narrative summary, potentially

[1] meta-analysis-and-systematic-review

including a meta-analysis if the data allows. On the other hand, meta-analysis (MA)[2][3] is a statistical technique within or alongside a systematic review that quantitatively pools data from similar studies to get a single, more precise estimate of an effect. It examines data from a number of independent studies addressing a common research question, in order to determine overall trends. It synthesizes quantitative data and computes a combined effect size and variance measures across all of the studies. By combining these effect sizes, the statistical power is improved and can resolve uncertainties or discrepancies found in individual studies. All MAs are part of an SR, but not all SRs contain a meta-analysis. A systematic review can conclude with a narrative synthesis if studies are too different (heterogeneous) to combine statistically, making a meta-analysis impossible. Thus, it can be said that an SR is a whole project (finding & summarizing), and an MA is a powerful tool (statistical combination) used in the summary stage, but not all SRs use it (Ahn & Kang, 2018; Kim, 2023).

SR and MA have numerous benefits[4] as offering a complete summary of existing research on a topic, identifying research gaps, consolidating evidence from diverse sources, explaining inconsistencies between study results (heterogeneity) through methods like meta-regression, uncovering new insights, guiding future research, and others. They synthesize existing research to provide more precise, reliable, and powerful conclusions than single studies, guiding evidence-based decisions and improving patient outcomes by reducing bias and clarifying knowledge. They offer a comprehensive overview, increase statistical power by combining data, and can uncover trends and generate new hypotheses, making them crucial for healthcare, policy, and science. MAs are an integral part in influencing health policies, supporting research grant proposals, and shaping treatment guidelines. They consolidate their role as a fundamental methodology in metascience. MA is often, but not always, an important component of a systematic review (Ahmad, Khan & Tiwari, 2024; Bangdiwala, 2024; Yuan & Hunt, 2009).

Due to the importance of SRs and MAs, a review of the literature revealed a plethora of studies that presented SRs and MAs of AI and translation, linguistics or education. The first group of studies focused on general AI translation tools and their performance in multilingual contexts. Ssemugabi (2025) reviewed the role of AI in modern language translation, highlighting its expanding societal applications and the rapid evolution of translation technologies. Yang (2025) conducted an SR of online AI translation tools used in English language learning and teaching, emphasizing their pedagogical value and limitations. Batubara et al. (2025) synthesized findings across diverse fields to examine how AI is applied in translation, identifying challenges, innovations, and broader impacts. Additional review and MA studies, such as Chan and Tang's (2024) SRs of GPT-based translation and Elsadig's (2024) review of AI's impact on language translation, further demonstrate the growing interest in evaluating AI translation quality, accuracy, and usability. Together these studies show that AI translation tools have been widely examined across global languages, with consistent attention to accuracy, error patterns, and technological development. Mohamed et al. (2024) reviewed studies on the impact of artificial intelligence on language translation. Nguyen et al. (2025) reviewed the benefits and challenges of AI translation tools in translation education at the tertiary level; Deng and Yu (2022) reviewed studies on machine-translation-assisted language learning for sustainable education.

A second group of studies explored AI translation in specialized or technical domains. For example, Sun and Han (2025) examined financial translation in the digital age, focusing on human–AI collaboration and the unique challenges posed by financial terminology. In medical domain, Mai et al. (2024) reviewed text-to-image translation algorithms in medicine, while McNaughton et al. (2023) synthesized research on medical image translation using machine learning. Noll et al. (2023) conducted a scoping review on machine translation (MT) of standardized medical terminology, highlighting the complexity of domain-specific vocabulary. Meta-analytic work by Teibowei and Mbete (2023) evaluated the efficacy of MT in biomedical texts, offering quantitative insights into translation performance. Low-resource languages have also received attention, with Yazar et al. (2023) and Tafa et al. (2025) reviewing neural MT challenges in languages with limited training data. Zappatore, & Ruggieri (2024) conducted a methodological multi-criteria review of studies that adopted an MT in the healthcare sector. These studies emphasized that AI translation performance varies significantly across domains, with terminology density, technical specificity, and resource availability shaping translation outcomes.

A third group of studies examined AI translation and AI-enabled tools within language learning and educational settings. Ali et al. (2025) conducted a meta-analysis on the acceptance and use of AI applications in education, identifying factors that influence adoption and effectiveness. Chen et al. (2025) provided a systematic review and meta-analysis of AI-enabled assessment in language learning, focusing on design, implementation, and learning outcomes. Lee (2023) synthesized evidence on the effectiveness of MT in foreign language education, offering quantitative estimates of its pedagogical impact. Mirzaeian and Oskoui (2023) reviewed the use of Google Translate in foreign language learning, highlighting both benefits and persistent challenges. Together, these studies demonstrate that AI translation tools are increasingly integrated into language learning environments,

[2] https://www.sciencedirect.com/topics/medicine-and-dentistry/meta-analysis
[3] https://en.wikipedia.org/wiki/Meta-analysis
[4] Gemini

where they support comprehension, vocabulary development, and writing, while also raising concerns about overreliance and accuracy.

Additionally, fewer SR studies focused specifically on Arabic MT. Omar and Salih (2024) conducted a systematic review of English–Arabic MT post-editing, identifying recurring error types and implications for translation pedagogy. AlGhamedi (2024) examined constraints affecting neural MT quality across language pairs, including Arabic, and discussed both human and automated evaluation methods. These studies highlight the challenges posed by Arabic diglossia, semantic ambiguity, morphological richness, and orthographic variation. However, these studies provide valuable insights into Arabic MT performance, the number of SR studies that exclusively focus on Arabic remains limited compared to other languages.

Although numerous SR and MA studies have examined AI translation tools, specialized translation domains, and AI-supported language learning, none have synthesized findings across a single researcher's comprehensive body of AI-related studies. Existing SR reviews tend to focus on specific language pairs, narrow domains, or particular applications of AI translation. No study has integrated evidence across phonology, terminology, metaphorical expressions, discourse, and translation performance within a unified analytical framework. Moreover, no SR or MA studies have examined AI performance across 20 studies conducted by the same researcher, offering a unique opportunity to analyze thematic patterns, consistency, and variability across a coherent research program. This gap underscores the need for the current SR and MA study. Given this gap, this study aims to conduct an SR and MA of 20 articles by the author published between 2024-2025 on the use of some AI models (Microsoft Copilot (MC), DeepSeek (DS) and Google Translate (GT) in translation, linguistics and education. It aims to answer the following question: What do Al-Jarf's 2024–2025 AI studies collectively reveal about AI performance, limitations, and sociolinguistic implications?

The present SR and MA are intentionally limited to studies published between 2024–2025, ensuring that differences in performance reflect linguistic factors rather than technological drift. During this period, the AI models examined across all 20 studies (MC, DS, GT) remained architecturally stable, with consistent training paradigms and evaluation mechanisms. This controlled window eliminates confounding variables such as model updates, training-data shifts, and algorithmic changes, allowing the analysis to isolate linguistic performance rather than technological noise.

A unique strength of this study is that all 20 articles evaluate the same AI models under similar conditions, using comparable tasks across multiple linguistic domains. This rare level of methodological alignment reduces variation in research design, datasets, metrics, user populations, and tool versions, resulting in a stable analytic environment that strengthens the validity of the effect sizes. The review is therefore a controlled performance synthesis, not a historical survey.

This is also the first SR and MA to synthesize a single researcher's entire AI research program across multiple linguistic domains using the same models within a controlled time frame. Unlike typical SR/MAs that combine unrelated studies from different researchers, eras, and methods, this study avoids such noise and achieves a level of coherence and internal validity that other reviews cannot.

The study integrates AI performance across diverse linguistic levels—translation, phonology, metaphor, terminology, discourse, cultural expressions, error analysis, and narrative generation—creating a multimodal linguistic synthesis unprecedented in the field. Because the same models were evaluated across tasks with consistent criteria, the MA provides a unified performance profile that reveals strengths, weaknesses, patterns, inconsistencies, and cross-task correlations.

Finally, the study introduces a new methodological model for AI evaluation: a longitudinal, multi-study, single-researcher synthesis of AI behavior. This is a new model for AI evaluation that demonstrates how a coherent AI evaluation program can be built, how comparable datasets can be generated, and how cross-study patterns can be meta-analytically integrated. The findings offer actionable insights for AI pedagogy, translation training, model development, curriculum design, and AI literacy research.

2. Methodology

2.1 Study Corpus

The present SR and MA study is based on 20 research articles by the author published between 2024 and 2025. They examine the performance of contemporary AI models (MC, DS, and GT) in Arabic linguistics, translation, and education. These studies evaluate how different AI systems translate, transliterate, interpret, or analyze linguistic items across languages. They compare performance, accuracy, and error types. Together they provide a rich dataset for quantitative and qualitative synthesis. The studies fall into seven thematic clusters, each representing a distinct dimension of AI linguistic behavior.

The largest group of studies focuses on AI translation of technical terms and metaphorical expressions. These articles examine how AI models handle culturally embedded expressions, specialized terminology, and figurative language. Topics include translation of

chemical compound names (Al-Jarf, 2022f), the translation of abu/umm animal and plant folk names (Al-Jarf, 2025j), brand names (Al-Jarf, 2025f), and folk medical terminology (Al-Jarf, 2025t), Gaza–Israel war terminology Al-Jarf, 2025b), medical terms (Al-Jarf, 2024e), zero expressions (Al-Jarf, 2025v), and "sleep" expressions (Al-Jarf, 2025u). Additional studies compare GT's performance in 2012 & 2025 (Al-Jarf, 2025l) and GT's translation of full-text Arabic research articles (Al-Jarf, 2025a), Copilot's translation of contrastive emphatic negation (Al-Jarf, 2025i), Arabic expressions of impossibility (Al-Jarf, 2025s), and Arabic grammatical terms used metaphorically (Al-Jarf, 2025k). Together, these studies provide extensive data on AI accuracy, error patterns, and semantic fidelity in translation.

A second cluster examines AI and phonological processing such as AI decoding of phonologically distorted or encrypted Arabic on social media (Al-Jarf, 2025e), AI transliteration of borrowed English nouns containing /g/ into Arabic ((Al-Jarf, 2025c), and pronunciation errors in Arabic content narrated by AI on YouTube (Al-Jarf, 2025o). These studies assess the extent to which AI models can interpret phoneme-to-grapheme correspondence, handle non-standard orthography, and produce intelligible speech, offering insight into AI's phonological competence.

A third group addresses AI's ability to recognize and decode Arabic, Japanese and Chinese calligraphic text images by Gemini, Ernie ViLG, and Google Translate (Al-Jarf, 2025d). It evaluates AI's ability to process stylized, non-standard, and visually complex Arabic, Japanese and Chinese scripts, highlighting challenges in multimodal text recognition.

A fourth thematic cluster explores how AI interacts with real-world discourse, political language, and encrypted communication. Examples include AI decoding of encrypted Arabic on Facebook and YouTube ((Al-Jarf, 2025e),) and Copilot's translation of emphatic negation in Arabic discourse (Al-Jarf, 2025i). These studies demonstrate how AI performs in context-rich, culturally loaded environments and contributes a sociolinguistic dimension to the review.

A fifth cluster compares AI and student translators in rendering Arabic expressions of impossibility (Al-Jarf, 2025s), as well as human versus AI translation of chemical compound names (Al-Jarf, 2022f). They contain quantitative accuracy measures, error counts, and proportions, and involve the same AI model (MC), enabling effect-size comparisons. Their translation performance makes them suitable for meta-analytic synthesis.

A sixth group examines the linguistic competence of 4 AI models and the different errors they make. This study investigates the types of linguistic questions MC, DS, Gemini, and Monica cannot answer accurately. It provides qualitative and quantitative error patterns that reveal structural limitations in AI linguistic knowledge and can be integrated into a mixed-methods synthesis.

Finally, a seventh cluster addresses human attitudes toward AI and academic practices. These studies examine editors' and publishers' views on AI-generated research (Al-Jarf, 2025r), and Arab instructors' views on AI-generated student assignments (Al-Jarf, 2024b). They include proportions (e.g., acceptance vs. rejection rates) and are suitable for meta-analysis of proportions. This cluster adds a human-factors perspective to the review, contextualizing AI performance within academic and pedagogical practices.

2.2 Eligibility (Inclusion & Exclusion) Criteria

To be included in the corpus, the studies must be authored by Reima Al-Jarf, have been published between 2024–2025, and should include extractable quantitative or qualitative data. An external database search was not required because the corpus is a closed, predefined research program consisting of all AI-related studies authored by Al-Jarf during 2024–2025. These studies form a complete, self-contained dataset, indexed in multiple platforms (Google Scholar, ResearchGate, Semantic Scholar, Academia, SSRN, and Scopus and others) and is publicly available across major academic platforms and represents the full scope of the author's AI research program.

Since the aim of this SR/MA is to synthesize the author's entire body of AI research within a controlled two-year period, the dataset is already exhaustive and no additional studies exist outside this corpus. Therefore, an external search would not yield new eligible studies and was methodologically unnecessary.

The author's studies on AI and translation authored before 2024 were excluded such as in Google's English-Arabic translation of technical terms (Al-Jarf, 2021a; Al-Jarf, 2016a); electronic translation between Arabic and European languages (Al-Jarf, 2012). Duplicate AI studies presented at conferences were also excluded. Similarly, the author's studies on students' performance in translating technical terms and formulaic expression were excluded such as expressions of impossibility (Al-Jarf, 2024a); numeral-based (Al-Jarf, 2023d); Ibn (son) and Bint (daughter) (Al-Jarf, 2023c); time metaphors (Al-Jarf, 2023f); dar (house) and bayt (home) (Al-Jarf, 2022a); common names of chemical compounds (Al-Jarf, 2022f); color-based (Al-Jarf, 2019b), and om- and abu-expressions (Al-Jarf, 2017a); binomials (Al-Jarf, 2016b); polysemes (Al-Jarf, 2022b); pedagogical implications for translating the Gaza-Israel war terminology (Al-Jarf, 2024c); whether Arabic and foreign shop names should be translated (Al-Jarf, 2024d); and

author's studies on translation technologies were excluded technology integration in translator training (Al-Jarf, 2017b) and how to use the OmegaT (Al-Jarf, 2009a) and

Likewise, the author's linguistic studies were excluded as: student-translators' difficulties with English word+preposition collocations (Al-Jarf, 2022h); Arabic and English loan words in Bahasa (Al-Jarf, 2021b); English neologisms (Al-Jarf, 2010b); English and Arabic plurals (Al-Jarf, 2020); interlingual pronoun errors (Al-Jarf, 2010a); word+particle collocations (Al-Jarf, 2009b); SVO word order errors (Al-Jarf, 2007); grammatical agreement errors (Al-Jarf, 2000); and multiple Arabic equivalents to English medical terms (Al-Jarf, 2018b); formation of hybrid compounds with foreign lexemes in Arabic (Al-Jarf, 2023b).

Moreover, the author's studies on pronunciation and transliteration were excluded as semantic and syntactic anomalies of Arabic-transliterated compound shop names (Al-Jarf, 2023e); deviant Arabic transliterations of foreign shop names and decoding problems among shoppers (Al-Jarf, 2022c); English transliteration of Arabic personal names with the definite article /al/ on Facebook (Al-Jarf, 2022d); gemination errors in Arabic-English transliteration of personal names on Facebook (Al-Jarf, 2022e); variant transliterations of the same Arabic personal names on Facebook (Al-Jarf, 2022k); difficulties with phoneme-grapheme relationships (Al-Jarf, 2019a); faulty consonant gemination in pronouncing English biomedical terms by Arab healthcare professionals (Al-Jarf, 2025w); splitting unsplittable foreign words in casual speech (Al-Jarf, 2025x); vowel pronunciation errors in English biomedical terminology by Arab Healthcare Professionals (Al-Jarf, 2025y); pronunciation errors in English silent consonants (Al-Jarf, 2025z); proper noun pronunciation inaccuracies (Al-Jarf, 2022g); mobile audiobooks for listening comprehension (Al-Jarf, 2021c); effect of background knowledge on auditory comprehension in interpreting courses (Al-Jarf, 2018a); the effects of listening comprehension and decoding skills on spelling achievement (Al-Jarf, 2005a); spelling, listening and decoding skills (Al-Jarf, 2005b); text-to-speech software for promoting decoding skills and pronunciation accuracy (Al-Jarf, 2022i); text-to-speech software as a resource for independent interpreting practice (Al-Jarf, 2022j); YouTube videos as a resource for self-regulated pronunciation practice (Al-Jarf, 2022l); TED talks as a listening resource (Al-Jarf, 2021d); clipping of borrowings in spoken Arabic (Al-Jarf, 2023a).

2.2 Corpus Characteristics

The diversity of the corpus is methodologically justified by the multidimensional nature of AI linguistic performance. The 20 studies intentionally examine different modalities - translation, phonology, orthography, calligraphy recognition, semantic interpretation, and user attitudes - because AI linguistic behavior cannot be captured through a single task type. Despite this variation, all studies aim to evaluate AI performance in Arabic–English linguistic contexts. To ensure methodological coherence, studies were organized into seven thematic clusters, with quantitative outcomes synthesized within clusters and qualitative findings integrated narratively.

2.2 Information Sources

All of the studies in the corpus are in the author's publication lists. They are centralized, publicly available, indexed across all major academic platforms, and consistent across profiles (Google Scholar, ResearchGate, Semantic Scholar, SSRN, Academia, Harvard Library, with two on Scopus). To control for intervening variables in a way that no other SR/MA has accomplished, all 20 studies were conducted between 2024 and 2025, using the same AI models (MC, DeepSeek, GT, and Gemini and Ernie ViLG in one study), under consistent methodological conditions. This allows the meta-analysis to isolate linguistic performance rather than technological drift. The time period was limited to 2024-2025 because long-span reviews such as Lee (2023), which cover 20 years of MT evolution, mix rule-based MT, statistical MT, phrase-based MT, early neural MT, and Transformer-based systems, producing unstable and incomparable effect sizes. The present study avoids this problem entirely. In this study, the AI models used are all neural and LLM and no major changes have taken place during those 2 years. The study also provides a unified performance profile of recent AI models. Because the author's studies evaluate the same models across different tasks using consistent criteria, the meta-analysis can reveal strengths, weaknesses, patterns, inconsistencies, and cross-task correlations. This produces a holistic performance map of AI translation and language behavior—something the field currently lacks.

2.3 Study Design

This study adopts an SR design following the PRISMA principles (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). The review synthesizes a closed, predefined corpus consisting of all AI-related studies authored by Al-Jarf in 2024–2025. Because the corpus is complete and bounded, the review process includes the standard PRISMA components of eligibility criteria, study selection, data extraction, and quality assessment. Where the included studies provide quantitative accuracy measures, error counts, or proportions, a meta-analysis is conducted to generate pooled effect sizes. For studies that report qualitative linguistic patterns or error types, a narrative synthesis is used. This mixed-methods design allows the review to integrate quantitative and qualitative findings into a unified analytical framework, consistent with PRISMA recommendations for complex linguistic datasets.

2.4 Data Extraction

From each study, the following data were extracted data: The AI model used (Copilot, DeepSeek, GT, Gemini, etc.), The sample size and description, the linguistic domain (metaphor, terminology, phonology, discourse, etc.) and context (e.g., social media, research articles, folk terms), analysis, research instrument such as surveys & content analysis, kind of task AI models were asked to perform (translation, phonology, calligraphy decoding, discourse, etc.), the results in terms of proportions and accuracy scores (% of correct and incorrect responses), error counts, and qualitative error types.

2.5 Quality Assessment

The included studies in the corpus were evaluated for methodologically consistent, clarity of outcome measures and comparability. They used the same AI models, and stable LLM architectures, were conducted in the same time window, they used similar evaluation criteria, they had clear outcome measures, and had no major methodological flaws. Because all 20 studies were conducted by the same researcher using stable LLM-based systems during 2024–2025, the risk of methodological heterogeneity was minimal.

2.6 Meta-analysis Procedures

The meta-analysis used proportion-based effect sizes in percentages, which is appropriate for studies reporting accuracy rates, error counts, recognition success, decoding success, and percentage agreement. A random-effects model was applied to pool results, and heterogeneity was assessed using the Q statistic and I^2 statistic. A mixed-methods synthesis was conducted: quantitative meta-analysis for translation accuracy and proportion-based outcomes, and qualitative synthesis for discourse-level and error-analysis studies. Subgroup analyses were performed where relevant (e.g., translation vs. phonology, metaphor vs. terminology, Copilot vs. DeepSeek). The calculations were carried out using manual computation, Excel, and SPSS, depending on the dataset. Because the 20 included studies were descriptive, accuracy-based, error-based, proportion-based, recognition-based, and decoding-based, proportions served as the unified effect size across all analyses, consistent with standards for meta-analyses of diagnostic accuracy and linguistic performance.

2.7 Data Synthesis

The synthesis combined quantitative and qualitative approaches. Studies reporting numerical outcomes (e.g., accuracy rates, error counts, or proportions of correct AI outputs) were summarized in Tables 1, 2, & 3 and descriptively and prepared for meta-analysis. Studies reporting qualitative linguistic patterns (e.g., error types, translation behaviors, or phonological deviations) were synthesized narratively and grouped into thematic clusters. Quantitative and qualitative findings were then integrated to provide a comprehensive profile of AI performance across the seven thematic domains.

2.8 PRISMA Flow Description

The number of records identified corresponds to the full set of 20 AI studies the author produced in 2024–2025. These records were obtained from the author's publication list and verified across her Google Scholar, ResearchGate, Academia, Semantic Scholar, SSRN, and Scopus profiles. Because the corpus is predefined and closed, all 20 records were screened. Each study was confirmed to be an AI study and all met the inclusion and eligibility criteria. Accordingly, all 20 studies were included in the final synthesis.

3. Results

3.1 Overview

This subsection presents the findings of 20 studies included in the review, organized into thematic clusters reflecting major domains of AI performance in Arabic and cross-linguistic contexts. The clusters cover phonology and pronunciation, calligraphy and orthography, linguistic reasoning, translation across multiple sub-domains, contrastive emphatic negation, encrypted Arabic, and comparative human–AI translation performance. Each cluster summarizes quantitative accuracy measures and qualitative error patterns reported in the original studies. The results are descriptive and reflect the outcomes of the included studies without interpretation or evaluation.

3.2 Study Characteristics

The 20 studies span a wide range of linguistic and multimodal tasks involving Arabic, English, Chinese, and Japanese. They examine the performance of generative AI systems (e.g., MC, DS, Gemini, Ernie ViLG) and neural machine translation tools (e.g., GT) across domains such as phonology, calligraphy recognition, metaphor interpretation, terminology translation, folk and cultural expressions, encrypted language, and full-text academic translation. Sample sizes vary from small sets of specialized terms (e.g., 60–100 items) to large corpora of expressions (e.g., 318 zero-expressions, 436 contrastive emphatic negation structures) and full research articles. Outcomes include quantitative accuracy rates, proportions of correct, partial, and faulty responses, and qualitative analyses of semantic, syntactic, pragmatic, and cultural error types. Together, the studies provide a diverse dataset capturing AI performance across multiple linguistic levels, genres, and communicative contexts.

3.3 AI Translation of Technical Terms and Metaphorical Expressions

There are 9 studies that share the same macro-domain: translation of medical terms, sleep terms, war terminology, zero-expressions, metaphorical grammar terms, folk medical terms, Abu-brand names, Abu/Umm metonymy, GT diachronic evaluation. They all belong to lexical translation, domain-specific terminology, semantic equivalence, AI translation accuracy. This cluster is subdivided into the following sub-clusters:

3.3.1 Sub-cluster 1: Domain-Specific Terminology:

This sub-cluster includes three studies examining AI translation performance on specialized terminology: (i) medical terms translated by MC and GT, (ii) English and Arabic sleep-related terms and formulaic expressions translated by MC and DS, and (iii) Gaza–Israel war terminology used in media and political discourse. Across 204 English and Arabic medical terms, GT produced correct equivalents for 74.5% of items. MC produced correct equivalents for 68.6% of items. Both systems performed better on Arabic → English than English → Arabic. Across 331 items (130 English sleep terms, 91 English formulaic expressions, 110 Arabic items), MC produced 91% correct for English sleep idioms, 79% correct for English formulaic expressions, and 48% correct for Arabic items. DS produced 91% correct for English sleep idioms, 71.5% correct for English formulaic expressions and 49% correct for Arabic items. Both AI systems performed better on English → Arabic than Arabic → English. For English → Arabic translation of Gaza-Israel war terminology, MC and GT produced identical equivalents for 38% of items. GT produced 22% correct equivalents. MC produced 20% correct equivalents. Both systems produced 18% correct equivalents with different wording. For Arabic → English translation, MC and GT produced identical equivalents for 58% of items. They produced 26% correct equivalents with different wording. Overall accuracy was higher in Arabic → English than English → Arabic.

Both MC and GT showed semantic errors, as literal translations of culturally or medically marked terms; transliteration of unfamiliar items; incorrect definiteness, word order, and derivational forms; contextual errors, such as misinterpreting نهجة as approach/method; inconsistent handling of polysemous terms, e.g., lupus used for both الثعلبة and الذئبة الحمراء. GT showed more literal and transliterated outputs, while MC occasionally produced explanatory equivalents. In translating sleep terms and formulaic expressions across both AI systems, literal, word-for-word translation was the dominant strategy; idiomatic expressions were frequently rendered literally (e.g., He drowned in sleep); DS provided more explanatory annotations (14%) than MC (3%); Arabic → English translation showed lower accuracy due to limited representation of Arabic idioms in training corpora; and nuance was often flattened, especially in metaphorical or culturally embedded expressions. In translating war terminology, the study documented difficulty translating weapons, military actions, war metaphors, and institutional terms; inconsistent equivalents across English and Arabic media sources; reliance on literal translation for metaphorical or politically charged expressions; and challenges in maintaining cohesion and conceptual accuracy in news translation tasks.

Across domain-specific terminology, AI systems showed moderate to high accuracy for transparent and well-represented terms, particularly in English → Arabic translation. Performance declined for idiomatic, metaphorical, culturally embedded, or politically charged terminology. Literal translation was the dominant strategy across all systems, with recurrent semantic, contextual, and syntactic errors. Arabic → English translation consistently showed lower accuracy due to the limited representation of Arabic idioms and specialized terminology in AI training corpora.

3.3.2 Sub-cluster 2: Structural and Formulaic Expressions

This sub-cluster includes two studies examining AI translation performance on structurally marked and metaphorical expressions: (i) zero-expressions translated by MC and GT, and (ii) Arabic grammatical terms (AGTs) used metaphorically, translated by MC, DS, and GT. Across 318 English and Arabic zero-expressions, 29% were correctly translated by both MC and GT, MC produced 52% noun + derived-adjective equivalents (e.g., صفرية); GT produced 50% noun + derived-adjective equivalents; MC produced 31% definite equivalents (e.g., التصنيف الصفري); GT produced 9% definite equivalents; GT produced 11% equivalents with awkward word order.; MC and GT produced 12% equivalents with reversed word order; 5% of outputs contained faulty derived forms from both systems. In translating Zero-expressions and across both AI systems: word-for-word translation was the dominant strategy; conceptual translation and modulation were rarely used; polysemous zero-expressions were frequently mistranslated; literal renderings produced semantically inaccurate equivalents (e.g., false zero → صفر زائف / صفر خاطئinstead of صفر غير حقيقي); idiomatic zero-expressions were mistranslated, including صفر على الشمال→ zero on the north (MC) / zero to the north (GT). Structural issues included reversed word order, awkward syntactic constructions, incorrect definiteness, and faulty derivational forms.

Across the 52 Metaphorical Arabic grammatical terms (AGTs), MC produced 43% correct translations; DS produced 29% correct translations; GT produced 23.5% correct translations; The three systems produced identical translations (correct or incorrect) for 57% of items. In translating Arabic grammatical terms used Metaphorically and across MC, DS, and GT, literal translation was the dominant strategy; metaphorical AGTs were often rendered literally, producing odd or humorous equivalents (e.g., بين بين → between between); polysemous AGTs were frequently mistranslated (e.g. المبتدأ and الخبر); culturally embedded expressions,

slogans, and expressions requiring historical background were difficult for all systems; MC produced the highest number of correct metaphorical equivalents, including: إن الحكاية فيها→ there is something fishy, بفعل فاعل→ by an unknown person, حاشا وكلا → absolutely not. DS produced some correct equivalents (e.g., بين بين→ in between), while GT showed the lowest accuracy and the highest literalness.

Across structural and metaphorical expressions, AI systems showed low accuracy relative to other terminology domains. Zero-expressions and metaphorical AGTs were predominantly translated word-for-word, resulting in semantic, syntactic, and contextual inaccuracies. Polysemy, idiomaticity, cultural content, and metaphorical usage posed consistent challenges. MC outperformed DS and GT on metaphorical AGTs, while both MC and GT showed similarly limited performance on zero-expressions.

3.3.3 Sub-cluster 3: Folk and Cultural Terminology

This sub-cluster includes three studies examining AI translation performance on culturally embedded Arabic expressions containing Om and Abu in folk medical terms, brand names, and animal/plant folk names. The studies compared MC and DS. In the translations of Folk medical terms with Om/Abu, and across 205 folk medical terms, DS produced 66% correct translations. MC produced 46% correct translations. MC produced 16% literal word-for-word translations; DS produced 11%. Both systems frequently translated أبو / أمliterally as mother and father. Both systems produced lexical variants (e.g., cerebral aneurysm for brain aneurysm) and equivalents with altered word order. In translating the Folk medical terms and across MC and DS, literal translation of أبو / أمwas common, producing outputs such as father of the knees for أبو الركب; idiomatic and culturally embedded terms were mistranslated; lexical variants and synonyms were used inconsistently; word order frequently differed from dictionary definitions and both systems struggled with specialized folk medical terminology.

In translating Abu-brand names, and across 100 brand names and three prompting tasks, MC produced literal word-for-word translations for 100% of items in all tasks. DS produced literal translations in Tasks 1 and 2, but in Task 3 (with product name added): DS transliterated 100% of brand names as proper nouns. 66% of DS transliterations were appropriate; 34% were inappropriate for grassroots brand names. MC and DS produced identical English equivalents for 83% of items in Tasks 1 and 2. DS produced double equivalents for 14.5% of items in Set 1. DS produced faulty annotations based on incorrect kunya-style inferences. In translating Abu-brand names and across tasks, MC consistently translated Abu literally as father of, regardless of prompt type or product association; DS shifted strategy only when product names were added, treating brand names as proper nouns; DS sometimes produced extraneous or incorrect annotations and grassroots brand names were frequently mistranslated due to misidentification of the intended referent.

In translating Denotative and metonymic Abu/Umm animal and plant names and across denotative and metonymic lists: in the Denotative names set, DS: 51% correct (no-domain prompt) and 51% correct (domain prompt); MC: 46% correct (no-domain) and 44% correct (domain); Both systems produced identical equivalents for 40% of items. In the Metonymic names, Both MC and DS produced fewer than 3% correct responses across all three prompts. MC produced 30% correct equivalents for Umm-names; 70% were faulty. DS produced 0% correct equivalents for Umm-names in the no-domain prompt and 97–99% faulty equivalents in the domain and metonymic prompts. In translating Animal and plant folk names and across MC and DS, denotative names were easier to translate, often rendered directly as the animal name; metonymic names were consistently misinterpreted as personal names; literal translation of Abu as father was common; faulty semantic guesses were frequent (e.g., assigning lizard as the referent animal for all metonymic items); DS often provided genus-level annotations rather than the specific animal or plant.

Across folk and cultural terminology, AI systems showed moderate accuracy for denotative folk names and some folk medical terms, but very low accuracy for metonymic expressions and grassroots brand names. Literal translation of أبو / أم, misclassification of metonymic names as personal names, faulty semantic guessing, and inconsistent handling of culturally embedded terms were recurrent patterns. DS outperformed MC on folk medical terms and denotative names, while both systems showed substantial limitations in metonymic and culturally grounded expressions.

3.3.4 Sub-cluster 4: Diachronic Machine Translation Performance

This subcluster includes one study comparing GT's performance in the Statistical Machine Translation (SMT) era (2012) and the Neural Machine Translation (NMT) era (2025). The study evaluated translations from Hungarian, German, Spanish, Turkish, and Japanese into both English and Arabic. In the SMT era (2012), translations from Hungarian, German, and Spanish into English were somewhat intelligible but generally literal, syntactically awkward, and lacking idiomatic nuance. Turkish and Japanese translations showed broken syntax, incoherence, and nonsensical phrasing. Semantic accuracy was low across all five languages. Arabic translations were particularly poor: lexical equivalents were inaccurate, sentence structure was jumbled and fragmented, and outputs were largely unintelligible. SMT performance was substantially weaker for typologically distant languages, especially Arabic, Turkish, and Japanese. By contrast, in the NMT era (2025), translations into both English and Arabic became intelligible, fluent, coherent, and stylistically natural. Syntax and word order were preserved, idiomatic and contextual meanings were handled

more accurately, and semantic, lexical, and syntactic accuracy improved across all five languages. The shift from SMT to NMT resulted in substantial gains in translation quality, with the most dramatic improvements observed in Arabic.

A second time-based pattern appears when looking at the twenty studies conducted across 2025. The studies from early 2025—especially those dealing with zero expressions, Gaza–Israel terminology, and metaphorical grammatical terms—showed lower accuracy and more hallucinations because these areas are culturally dense, pragmatically complex, and not well represented in training data. The AI models used at that time had also not yet received the mid-year refinements. In contrast, studies from the second half of 2025 showed higher accuracy, particularly in standardized and scientific domains, reflecting both easier content and the gradual improvements that AI systems typically undergo throughout the year. This difference does not represent a shift from "old" to "new" models, but rather the normal, ongoing refinement of learning-based architectures as training data expands and tuning improves. Taken together, the two diachronic comparisons—SMT vs. NMT (2012 vs. 2025) and early vs. late 2025—show that AI translation performance improves both across years and within the same year, shaped by technological updates and by the difficulty of the linguistic domain.

3.4 AI Translation of Contrastive Emphatic Negation in Arabic discourse

This cluster includes one study that evaluates MC's English translation of 436 Arabic contrastive emphatic negation (CEN) expressions containing a range of negative particles and structures. Across the 436 CEN expressions, MC produced correct translations for 65% of the items. Accuracy was higher for expressions containing a single negative particle than correlative negation (e.g., لا ... ولا "neither ... nor"). MC performed better when the negation was literal, the structure was transparent, the expression was short and syntactically simple. The study identified several recurrent error types in MC's translations: (i) Semantic errors as failure to convey idiomatic or culturally embedded meanings, difficulty with polysemous words, and literal translation of fixed expressions. (ii) Syntactic errors as faulty structure, incorrect wording, inappropriate use of articles, incorrect selection of negative particles, mismatches in part of speech or derivational form. MC showed consistent difficulty when CEN expressions were Idiomatic, culturally loaded, pragmatically marked, metaphorical, and dependent on multiword expression recognition. The study also noted that MC tended to translate word-for-word rather than select the fixed English equivalent. CEN expressions were easier for MC than zero-expressions, Gaza–Israel war terminology, metaphorical grammatical terms, expressions of impossibility, folk medical terms with abu/om, and metonymic abu/umm animal and plant names.

3.5 AI Translation of Full-Text Academic Discourse

This cluster includes one study evaluating AI performance in translating full-text Arabic research articles into English, with a specific focus on educational polysemes (Al-Jarf, 2025). The study identified multiple instances in which AI systems mistranslated Arabic polysemous terms that carry both general and specialized meanings. Errors occurred across a wide range of educational terminology, including: صدق→ translated as *honesty* instead of *validity;* التحكيم / المحكمون→ translated as *arbitration* / *arbitrators* instead of *peer reviewing* / *reviewers;* المنهج المحوري→ translated as *axial* rather than *spiral* curriculum; رسالة→ frequently rendered as *message* instead of *thesis;* تصورات→ translated as *visions* instead of *models;* خطة→ translated as *plan* instead of *proposal;* لجنة المناقشة→ translated as *discussion committee* instead of *defense committee*. Additional mistranslations were observed in terms related to school grade levels, instructional guides, academic content, and teaching load. Across the full-text translations, AI systems showed consistent tendencies: (i) Literal rather than conceptual translation, especially for terms with domain-specific meanings in education. (ii) Failure to distinguish between general and technical senses of polysemous words. (iii) Incorrect selection among multiple English equivalents, particularly when Arabic terms had one-to-many mappings. (iv) Context-insensitive rendering of terminology embedded in academic discourse. (v) Inaccurate translation of institutional and procedural terms, such as committees, curricula, and academic processes. These patterns indicate difficulty in maintaining semantic precision across extended academic texts. Overall, the study demonstrates that AI systems struggle with translating polysemous educational terminology in full-text Arabic research articles. Errors were systematic and centered on literal translation, mis-selection among multiple English equivalents, and insufficient sensitivity to academic context.

3.6 Comparison of AI and humans in Translating Chemical compound names and expressions of impossibility

This cluster includes two studies comparing the translation performance of MC with that of undergraduate student-translators. The studies examined (1) translation of common names of chemical compounds by AI and student translators (Al-Jarf, 2025m), and (2) translation of English and Arabic expressions of impossibility (Al-Jarf, 2025). In translating Chemical compound names, MC produced 72% correct translations across the sample. 40% correct in Arabic → English & 32% correct in English → Arabic. Student-translators in the earlier study produced fewer than 20% correct translations and left 55% of items blank. MC produced 52% correct translations in the Expressions of impossibility sample. Student-translators produced fewer than 35% correct translations and left many items blank. Both MC and students performed better in Arabic → English than English → Arabic.

Across both studies, MC showed strong performance on transparent, literal, or structurally simple items, frequent word-for-word translations, especially for idiomatic or metaphorical expressions, occasional faulty lexical choices, incorrect derivatives, or literal

renderings of culturally marked expressions, difficulty with opaque idioms, polysemous terms, and metaphorical EIs and tendency to provide only one equivalent unless prompted for all possible translations. Examples included: Stearic acid → حمض الستيريك (transliteration instead of حمض الشمع) & once in a blue moon → مرة واحدة في القمر الأزرق (literal, culturally awkward). By contrast, student-translators showed limited domain knowledge (chemical terminology, idioms, proverbs), reliance on literal translation and transliteration, difficulty with opaque expressions, unfamiliar vocabulary, cultural references, frequent omissions, and use of paraphrase/explanation as a compensatory strategy. Students' errors were attributed to limited linguistic mastery, insufficient exposure to idioms, and lack of background knowledge.

Together, the two studies show that MC consistently outperforms student-translators in both technical terminology and idiomatic expressions, though both rely heavily on literal translation. MC demonstrates higher accuracy and broader lexical access, while students show greater difficulty with unfamiliar or culturally embedded expressions. Both groups perform better in Arabic → English translation than in English → Arabic.

3.7 Phonology Cluster

This cluster includes two studies: (1) transliteration of English nouns containing /g/ into Arabic (Al-Jarf, 2025), and (2) pronunciation accuracy in AI-narrated Arabic YouTube videos (Al-Jarf, 2025). Across 140 English nouns containing /g/, MC and GT produced highly similar Arabic transliterations, closely matching human transliteration patterns. Use of غ: 61% of items غ pronounced /ɣ/: 17% & غ pronounced /g/: 44%.; Use of ج: 37% of items ج pronounced /g/: 27% & ج pronounced /dʒ/: 10%; and Use of ق: extremely rare: MC: 1 item & GT: 1 item. AI systems produced one transliteration per word, whereas human transliterators produced multiple variants for 16% of the sample.

In the Arabic videos showed that AI narration has a natural voice quality, intonation, and expression, no grammatical or syntactic errors, frequent pronunciation errors, especially in short vowel diacritics, homographs, suffixes involving تاء التأنيث (ta/ti/tu), distinguishing person, gender, and tense. These errors affected intelligibility for both native and non-native listeners.

Across both studies, phonological errors were systematic rather than random: AI struggled with context-dependent pronunciation, especially when short vowels were absent. Homographs were frequently mispronounced due to failure to infer the correct diacritic from context. In transliteration, AI followed dominant regional conventions (غ in Levantine patterns, ج in Egyptian patterns) but did not reproduce the full range of human variation. AI rarely used ق, despite its presence in Saudi Arabic transliteration practices. In narration, AI mispronounced verb forms and suffixes that require morphological awareness (e.g., كتبتِ / كتبتُ / كتبتَ). Both studies highlight limitations in AI's ability to map orthography to phonology when Arabic relies on unvowelled text. Together, the two phonology studies show that AI systems perform well in consistency and surface-level phonological mapping, but struggle with context-dependent pronunciation, diacritic inference, and dialect-sensitive transliteration choices. Errors cluster around homographs, short vowels, and morphologically marked suffixes.

3.8 AI Interpretation of Encrypted Arabic in Spoken and Written Media

This cluster includes one study evaluating whether AI systems can interpret encrypted Arabic words and phrases used by social-media creators on YouTube (spoken) and Facebook (written) to evade algorithmic moderation (Al-Jarf, 2025). The study compared the performance of MC, DS, and GT. Across 74 encrypted political words and phrases: MC gave 56% correct interpretations, 16% partial interpretations and 27% faulty responses. DS gave 41% correct, 35% partial, 24% faulty. MC and DS identical correct responses: 36% of items Encrypted COVID-19 content (Facebook, written). Across 20 encrypted COVID-19 terms: MC rendered 60% correct, 25% literal translations, 10% partial, 5% omissions. DS rendered 50% correct, 15% literal, 5% partial, and 30% faulty. MC and DS identical correct responses: 35% of items. GT produced 42% word-for-word literal translations, 44.5% transliterations and 0% contextual interpretations. GT did not infer encrypted, distorted, slang, satirical, or parodic meanings. Examples included literal renderings such as القبة الزجاجية → *the glass dome* and *16□□□□* → *Viva 16* instead of *F-16*. Across both spoken and written encrypted samples: MC and DS demonstrated the ability to infer underlying meanings in some cases, but also produced partial or faulty interpretations. GT consistently provided surface-level outputs, limited to direct translation or transliteration. None of the systems reliably interpreted distorted, slang, or intentionally obfuscated forms. Generative models (MC, DS) showed partial contextual inference, whereas GT operated strictly at the lexical level. Overall, the study shows that generative AI systems outperform NMT in interpreting encrypted Arabic used to evade moderation, but accuracy remains limited. MC and DS provided partial contextual understanding, while GT produced literal or transliterated outputs without recognizing hidden or obfuscated meanings.

3.9 AI Recognition & decoding of Calligraphic Text Cluster

The calligraphy recognition cluster included one study that evaluated Gemini, Ernie ViLG, and GT's ability to recognize calligraphic text images in Arabic, Japanese, and Chinese. Across 15 Arabic, 7 Japanese, and 10 Chinese calligraphic samples, the three models showed markedly different levels of recognition accuracy: Gemini correctly matched 12/15 Arabic, 7/7 Japanese, and 9/10 Chinese

images. Ernie ViLG produced 0/15 correct Arabic matches, and correctly translated 4/7 Japanese and 3/10 Chinese images. GT failed to recognize all Arabic samples and generated partial or incoherent outputs for Japanese and Chinese calligraphy. These results indicate substantial variation in model performance depending on script type and visual complexity. It identified consistent error patterns across AI models: Gemini relied on multilingual training and semantic retrieval, enabling it to associate stylized calligraphic texts with known verses or idioms. Ernie ViLG frequently defaulted to culturally common themes when unable to decode the strokes, reflecting reliance on cultural priors rather than visual recognition. GT struggled with all stylized scripts due to its OCR pipeline being optimized for printed or clean handwritten text, leading to fragmented or incoherent outputs. Across all three languages, errors were driven by difficulties with distorted or ornamental strokes, non-linear writing paths, stylized ligatures and ambiguous character shapes. Overall, the calligraphy study demonstrates that multimodal AI systems vary widely in their ability to recognize and decode stylized scripts. Gemini showed the highest recognition accuracy, while Ernie ViLG and GT showed limitations, particularly with Arabic calligraphy.

3.10 AI's Linguistic Reasoning and Question-Answering Accuracy

This cluster includes one study examining AI performance on 50 specific linguistic and translation questions posed to five AI systems (MC, DS, GT, Gemini, and Monica) between 2023 and 2025 (Al-Jarf, 2025). The 50 questions covered a wide range of linguistic domains: phonology, transcription, morphology, lexical questions, pragmatics and culture, explanation or translation of Arabic grammatical terms, identification of books, classical Arabic storytelling, handwritten-to-typed text conversion, translation of technical terms, metaphorical expressions and metonyms, bibliographic and scholarly workflow tasks. Across these categories, all five AI systems revealed inaccuracies, with errors appearing in every domain in the sample. AI shortcomings included: (i) frequent errors in sound–symbol associations and transcription accuracy. (ii) incorrect analyses of word structure and inaccurate lexical choices. (iii) misinterpretations of culturally embedded expressions and context-dependent meanings. (iv) faulty or incomplete explanations and mistranslations of technical grammatical terms. (v) inability to recognize certain titles or provide accurate bibliographic information. (vi) Telling stories from classical Arabic lierature. (vii) failure to convert handwritten Arabic into typed text. (viii) Technical terminology: inconsistent or incorrect translations of specialized terms. (xiv) Metaphors and metonyms: frequent literal interpretations and mistranslations and (xv) fabricated references and incomplete organization of citation data. Across all categories, errors reflected surface-level processing, limited contextual inference, and insufficient cultural grounding. Overall, the study shows that AI systems struggle with fine-grained linguistic reasoning, context-dependent interpretation, and tasks requiring cultural or scholarly depth. Errors were systematic across phonological, morphological, lexical, pragmatic, and bibliographic domains, indicating persistent limitations in AI's ability to answer specific linguistic questions accurately.

3.11 Human Attitudes Toward AI-Generated Academic Work

This cluster includes two studies examining human evaluations of fully AI-generated academic outputs: (i) Editors and publishers' attitudes towards the publication of AI-Generated Research Articles in Scholarly Journals (Al-Jarf, 2025), (ii) Arab Instructors' attitudes towards students' assignments and research papers generated by AI (Al-Jarf, 2024b). Across both studies, rejection of fully AI-generated academic work was overwhelmingly high: Editors and publishers: 90% rejected the publication of fully AI-generated research articles in scholarly journals. University instructors: 98% did not accept AI-generated assignments or research papers submitted by students. These proportions indicate a strong consensus against accepting fully AI-generated academic outputs in both professional and educational contexts. Across the two studies, participants provided consistent reasons for rejecting fully AI-generated work: (i) Integrity and ethics: concerns about cheating, dishonesty, plagiarism, and lack of academic integrity. (ii) Authorship and responsibility: belief that human authors must be accountable for ideas, analysis, interpretation, and conclusions. (iii) Quality and credibility: doubts about the accuracy, coherence, and reliability of AI-generated content. (iv) Skill development: fear that students would not acquire research, writing, or critical thinking skills if they relied on AI. (v) Verification challenges: concern that students lack the expertise to detect AI-generated errors, mistranslations, or fabricated references. (vi) Fairness: risk of unfair evaluation when some students submit AI-generated work and others do not. Despite rejecting fully AI-generated outputs, both groups expressed openness to partial or supportive uses of AI, including: brainstorming, literature review assistance, data analysis and visualization, summarization, translation, editing and proofreading. Both groups emphasized the need for declaring AI use, verifying and correcting AI outputs, and maintaining human responsibility for meaning, synthesis, and interpretation. Instructors additionally recommended raising students' awareness of institutional policies regarding AI use. Together, these studies show a strong and consistent pattern: editors, publishers, and instructors overwhelmingly reject fully AI-generated academic work but accept AI as a supportive tool when human authors remain responsible for the intellectual contribution and accuracy of the final product.

4. Discussion

4.1 Difficulty Level Across Translation Domains and AI Models

MC was used in 15 studies, DS in 8 and GT in 7 studies. The % of correct translation equivalents to the term and metaphorical expressions in 12 specialized domains are shown in Table 1.

Table 1: Summary of Correct Translation Percentages Across All AI Models and Translation Domains

Study / Domain	Sub-domains	MC % Correct responses	DS % Correct responses	GT % Correct responses	Ernie % Correct responses	Gemini % Correct responses
Sleep terms & metaphorical expressions	English terms	91%	91%	–	–	–
	English Formulaic Expressions	79%	71.5%	–	–	–
	Arabic terms and formulaic expressions	48%	49%	–	–	–
Chemical Compound Names	-	72%	–	–	–	–
Contrastive Emphatic Negation	-	71%	–	–	–	–
Medical Terms	-	68.6%	–	74.5%	–	–
Expressions of Impossibility	-	52%	–	–	–	–
Om/Abu Folk Medical Terms	-	46%	66%	–	–	–
Abu/Umm Animal & Plant Names	Abu- with (denotative name)	46%	51%			
	Umm name	30%	1%			
	abu/umm with metonymic name	15%	1%	–	–	–
Grammatical terms used metaphorically	-	43%	29%	23.5%	–	–
Zero-Expressions	-	29%	–	29%	–	–
Gaza–Israel War Terminology	English → Arabic	20%	–	22%	–	–
	Arabic → English	29%	–	23%	–	–
Abu-Brand Names	abu-brand names (no specification prompt)	0%	0%	–	–	–
	abu-brand names (brand names only prompt)	0%	0%	–	–	–
	abu-brand names (brand name + product name prompt)	0%	66%	–	–	–
Calligraphic text images	Arabic	–	–	0%	0%	80%
	Japanese	–	–	0%	57%	100%
	Chines	–	–	0%	30%	90%

Table 1 shows that in translating English sleep terminology, sleep metaphorical expressions, Chemical Compound Names, Contrastive Emphatic Negation, & Medical Terms and These domains share the opposite properties, and the AI models achieved 70–90% Accuracy. This is because these types of terms and metaphorical expressions appear in bilingual dictionaries, medical corpora, Wikipedia, scientific texts, translation memory datasets. They have transparent semantics. Their meaning is: literal, stable, domain-specific, and not culturally embedded. Example: triglycerides → الدهون الثلاثيةShoebill → أبو مركوب. They also have a strong cross-linguistic equivalence. Medical and scientific terms often have one-to-one mappings, standardized terminology, and international usage which reduces ambiguity. Additionally, they have a better morphological alignment. Terms like *diabetic ketoacidosis* & *cavernous carotid aneurysm* have predictable structures that AI can parse.

Some domains sit in the middle (50–70%). These are from "semi-transparent" domains such as contrastive, emphatic negative structures (71%), expressions of impossibility (52%), folk medical terms (46–66%) and Gaza–Israel AR → EN (58% identical). They combine some literal structure, some cultural content, some polysemy, and some domain knowledge. AI can partially succeed, but not consistently.

Domains with below 50% correct responses are systematically difficult for AI models to translate. Across all 20 studies, the domains with less that 50% correct responses share deep structural properties: High cultural loading as in Abu/Umm metonymic names,

folk expressions, zero-expressions, Arabic grammatical terms used metaphorically, Gaza–Israel terminology (EN→AR). These require cultural grounding, not just linguistic mapping. Structures in these domains are characterized by polysemy + metaphor + context dependence. AI models struggle when the surface form is misleading, the meaning is not compositional, the expression is idiomatic or metaphorical, and the referent is not explicitly stated. Examples: أبو الرّكّب , المبتدأ ونحن الخبر , الشمال صفر على ,أبو الشّبّت, *zero for zero approach*. Moreover, these structures have sparse representation in training data. These expressions rarely appear in English-dominant corpora, parallel translation datasets, and formal written Arabic. So the model has no statistical memory of them. Finally, AI models lack world knowledge or encyclopedic grounding. Metonymic names require knowing which animal is associated with which folk name, which plant is implied, which disease is referred to, and AI models cannot infer this without explicit training.

DeepSeek exhibited a systematic fallback behavior in the metonymic prompt, defaulting to 'lizard' as the referent animal for all items regardless of the folk name's actual meaning and even when the surface form gave no indication of a reptile. This pattern indicates a fallback hallucination pattern rather than a translation strategy and a lack of cultural grounding and an inability to map metonymic Abu-names to their intended referents.,

4.2 *Why Is Arabic → English Easier Than English → Arabic Translation*

In these studies, Arabic → English is easier than English → Arabic for the AI models across the entire dataset. This is attributed to Training data bias. AI models are trained on English-dominant corpora, English-aligned parallel data, English-centric web content. So they "understand" English better because English has simpler morphology, but Arabic requires definiteness, gender, number, case, and derivational patterns which English does not. Arabic equivalents require more lexical decisions. Example: zero fraction → صفر الكسر or كسر الصفر; *fibroglandular tissues* → multiple possible word orders. English equivalents are usually simpler. Arabic idioms also are underrepresented in corpora. AI models rarely see: folk expressions, metonymic names, Abu/Umm structures, metaphorical AGTs. So Arabic→English is easier because the model is decoding, not encoding.

4.3 *Why The Number of Items Is Not a Factor in AI Translation Accuracy*

Results show that across all 20 studies, the size of the dataset (number of items) had no meaningful correlation with AI translation accuracy. Findings show that difficulty is linguistic, not numerical. Across your studies, the number of items ranged from 20 items (chemical compounds), 40–60 items (AGTs, impossibility expressions), 100 items (Abu-brand names), 205 items (folk medical terms), 318 items (zero-expressions), 436 items (contrastive emphatic negative structures). Yet accuracy did not follow the size of the dataset. Results show that small datasets sometimes had high accuracy as Chemical compounds (60 items): 72% and English idioms (40 items): 91%. Large datasets sometimes had low accuracy as Zero-expressions (318 items): 29%, CEN (436 items): 71% and Folk medical terms (205 items): 46%. Medium-sized datasets also varied widely as AGTs (40 items): 23.5–43%, Abu-brand names (100 items): 0–66% and Denotative Abu/Umm names (60 items): 46–51%. There is no pattern linking dataset size to performance. The Real Factors Are Linguistic, Not Quantitative The global synthesis shows that accuracy depends on: (i) Semantic transparency: Literal → high accuracy Metonymic → low accuracy, (ii) Cultural grounding: Folk names → low Scientific terms → high, (iii) Domain stability: Medical terminology → stable War terminology → unstable, (iv) Polysemy and metaphor: Arabic grammatical terms used metaphorically → low Idioms → high (because they are frequent in corpora), (v) Corpus representation: Common expressions → high Rare expressions → low, (vi) Direction of translation: Arabic→ English → higher English → Arabic → lower, (vii) Model behavior: MC → literal bias DS → domain-sensitive but inconsistent GT → strong in standardized domains. None of these are affected by the sample size.

Meta-Conclusion

Across all 20 studies: High-accuracy domains are literal, frequent, standardized, scientific; medium-accuracy domains are semi-literal, semi-cultural, mixed transparency; and low-accuracy domains are idiomatic, cultural, metonymic, polysemous, context-dependent. Arabic → English consistently outperforms English → Arabic translations. Literal translation is the default strategy across all models. Cultural grounding is the single biggest weakness. Metonymy is the hardest domain for all AI systems. GT shows dramatic improvement over time (2012→2025). DS excels in folk medical terms and MC excels in idioms and structured expressions.

4.4 *AI vs Human Translation*

Table 2: % of Correct Translations by AI and Student-Translators

Study domain		% of AI Correct Responses	% of Students' Correct Responses
Common names of chemical compounds	Correct Translation equivalents	72%	20%
	Blank responses	0%	55%

Expressions of impossibility	Correct Translation equivalents	52%	35%
	Blank responses	0%	47.5%

The two studies in Table 2 show that AI translation accuracy surpassed that of student-translators. This is not because AI is "better," but because the task type favored AI's strengths and exposed students' weaknesses. The comparison between AI and studet-translators reveals the following differences:

(i) AI's breadth of knowledge vs. students' knowledge gaps. In both domains—chemical compound names and expressions of impossibility—AI models benefited from massive multilingual corpora, exposure to scientific terminology, access to standardized nomenclature, and familiarity with fixed expressions. By contrast, Student-translators struggled because they lacked background knowledge, were unsure of technical terminology, were unfamiliar with idiomatic impossibility structures, and hesitated to answer. This explains why AI produced 0% blank responses while students produced 47–55% blank responses.

(ii) The nature of the task favored AI's strengths. Both studies involved highly lexicalized items—chemical compounds and fixed impossibility expressions. These domains have stable meanings, standardized terminology, and widely available equivalents. Translation is more "lookup-based" than inferential. AI excels in this environment because it relies on pattern matching, statistical memory, and exposure to large bilingual datasets. Students rely on classroom knowledge, limited exposure, and personal experience, so the task type naturally amplifies AI's advantage.

(iii) Students' errors reveal cognitive load, not lack of ability. The high percentage of blank responses among students does not indicate weak translation ability. Instead, it reflects uncertainty, fear of giving incorrect answers, lack of domain knowledge, and limited training in these specific sub-fields. By contrast, AI has no hesitation and always produces an answer, even if imperfect. This behavioural difference explains the accuracy gap.

Meta-Interpretation.

These findings do not contradict the broader literature claiming that human translators outperform AI. They simply show that in highly technical, standardized domains, AI can outperform students, whereas in culturally loaded, metaphorical, or context-dependent domains, humans outperform AI by a wide margin. The dataset confirms a consistent pattern: AI is strong in scientific, lexical, standardized domains, whereas humans are strong in cultural, metaphorical, and inferential domains.

4.5 *Instructors and Editors' Reaction to AI-generated Content*

Tables 3 and 4 show that the majority of college instructors and journal editors reject AI-generated academic work for the same structural reasons. The problems are not moral or ethical, rather they are linguistic and epistemic. The weaknesses highlighted by instructors and editors align with the findings from the study on 50 types of linguistic questions that AI models consistently fail to answer accurately (Al-Jarf, 2025q). These include phonological and transcription errors, morphological and lexical inaccuracies, pragmatic and cultural misinterpretations, and faulty explanations of Arabic grammatical terms. Technical terminology, metaphorical expressions, and metonyms are frequently mistranslated, while bibliographic and scholarly workflow tasks often reveal fabricated references and gaps in organizing citations. Collectively, these errors underscore AI's tendency toward surface-level processing at the expense of linguistic depth, cultural fidelity, and conceptual precision. The reasons given by instructors and editors converge because the underlying limitations are the same: lack of cultural grounding, lack of deep reasoning, insufficient domain knowledge, hallucination under uncertainty, inability to maintain global coherence, and over-reliance on surface patterns.

The rejection patterns in Table 4 mirror the same structural weaknesses observed across the 20 translation studies, indicating that AI's limitations are consistent across both linguistic and academic tasks.

Table 3: % of Instructors and Editors Who Reject and Accept AI-Generated Academic Work

	% Reject	% Accept
Teachers (Assignments)	98%	2%
Journal Editors (Research Articles)	90%	10%

Table 4: Common Reasons Rejecting AI-Generated Academic Work by Teachers and Journal Editors

Reason for Rejection	Teachers (Assignments)	Journal Editors (Research Articles)	Reasons Given

Lack of originality	This is generic / template-like.	The argument lacks novelty.	AI produces safe, average, pattern-based text.
Hallucinated facts	Incorrect examples / invented citations.	References do not exist / fabricated data.	AI fills gaps with plausible but false information.
Inconsistent logic	Paragraphs don't connect.	Methodology is incoherent.	AI cannot maintain long-range argumentative structure.
Surface-level analysis	Too descriptive, no critical thinking.	No theoretical contribution.	AI summarizes but does not infer or critique.
Repetitive phrasing	Same sentence structure throughout.	Redundant wording.	AI optimizes for fluency, not stylistic variation.
Incorrect discipline- specific terminology	Misused technical terms.	Terminology inconsistent with field standards.	AI lacks deep domain grounding.
Ethical concerns	Student did not do the work.	Authorship integrity compromised.	Academic honesty and accountability.
Missing personal voice	Doesn't sound like the student.	Lacks scholarly voice.	AI cannot replicate authentic intellectual identity.
Overly confident tone	Sounds unnatural / too perfect.	Unsupported claims stated as facts.	AI defaults to assertive, polished phrasing.
Citation problems	Wrong citation style.	Inaccurate or incomplete references.	AI struggles with citation precision.

4.6 *Temporal Patterns in AI Performance*

A diachronic comparison of the twenty studies reveals a meaningful temporal pattern. Studies conducted in early 2025—particularly those involving zero expressions, Gaza–Israel terminology, and metaphorical grammatical terms—showed lower accuracy and higher hallucination rates. These domains are culturally dense and pragmatically complex, and the models used at that time had not yet reached the refinement levels observed later in the year. By contrast, studies conducted in the second half of 2025 demonstrated higher accuracy, especially in standardized and scientific domains. This improvement reflects a combination of domain accessibility and the gradual refinement characteristic of learning-based AI systems. The difference does not indicate a shift between "older" and "newer" models as in the GT translations in 20012 and 2025, but rather the evolving nature of AI architectures and the impact of training data expansion. Together, these findings highlight that AI translation performance is not static, but temporally sensitive—even within the same calendar year.

4.7 *Explanations for AI Strengths and Weaknesses*

The global patterns observed across the twenty studies can be explained by structural properties of current AI models. First, training data bias plays a central role: AI systems are trained primarily on English-dominant corpora, which means they have extensive exposure to English scientific terminology, idioms, and standardized expressions, but far less exposure to culturally embedded Arabic expressions. This imbalance explains why models perform well in literal, scientific, and high-frequency domains but struggle in culturally specific or low-resource areas. Second, inadequate and insufficient Arabic idiom representation in training data limits the models' ability to interpret metaphorical, folk, and formulaic expressions, leading to literal translations that miss the intended meaning. Third, AI models have persistent difficulty with polysemy, especially when the surface form is misleading or when the intended meaning depends on cultural knowledge rather than lexical cues.

A major weakness across all models is the lack of cultural grounding. AI systems do not possess encyclopedic knowledge of folk naming systems, metonymic Abu/Umm structures, or culturally anchored metaphors, which results in systematic misinterpretations and fallback hallucinations. Additionally, most models default to literal translation strategies, especially when uncertain, because literal mapping is statistically safer in the absence of contextual grounding. Weak performance in Arabic also reflects limitations in morphological modeling: Arabic's rich inflectional system, derivational patterns, and gender/number marking require deeper structural understanding than English, and current models often fail to parse these features accurately. Finally, AI systems lack pragmatic inference, meaning they cannot infer implied meanings, speaker intent, or culturally shared knowledge. This absence of pragmatic reasoning explains why models struggle with impossibility expressions, metaphorical grammatical terms, and politically loaded terminology.

4.8 *Cross-Cutting Insights*

Across all twenty studies, several cross-cutting patterns emerge that clarify both the strengths and limitations of current AI systems. On the strength side, AI models offer speed, accessibility, and broad exposure, making them valuable tools for learners and translators working with standardized or scientific terminology. Their ability to generate immediate outputs provides students with rapid feedback and supports vocabulary expansion in well-represented domains. However, the weaknesses are equally consistent. AI systems struggle with cultural nuance, idiomatic translation, phonetic accuracy, and the interpretation of metonymic or context-dependent expressions. These limitations reflect deeper issues in phonetic modelling, cultural grounding, and pragmatic inference, particularly in Arabic, where meaning often depends on morphology, discourse context, and shared cultural knowledge. These findings also point to clear opportunities for improvement. Future systems would benefit from integrating linguistic theory, developing Arabic-specific NLP models, and adopting hybrid human–AI approaches that combine computational efficiency with human cultural insight. Expanding Arabic-specific corpora—especially in folk, metaphorical, and culturally rich domains—would further enhance model performance and reduce systematic biases.

4.9 *Implications for AI Development & Translation Pedagogy*

The meta-analysis of the 20 studies herein carries broader implications for several fields: In translation pedagogy, the results show that students need more exposure to specialized terminology, training in inferential reasoning, practice with culturally loaded expressions, and strategies for avoiding blank responses. For dictionary building, the low accuracy in metonymic and folk domains underscores the need for richer lexicographic documentation of abu/umm names, zero-expressions, and metaphorical grammatical terms. The results also inform Arabic NLP research by identifying domains where models lack cultural grounding and require improved training data. For students and instructors, the findings clarify where human strengths lie and where AI can serve as a supplementary tool. Finally, the patterns observed in AI hallucinations have implications for media literacy and misinformation moderation, particularly in encrypted Arabic clusters where misinterpretations can spread rapidly. AI needs better cultural grounding, training on low-resource Arabic domains, mechanisms to detect uncertainty, better metonymy and metaphor handling, and improved Arabic morphological modelling.

4.10 *Limitations of The Meta-Analysis*

This meta-analysis study of 20 studies herein has the following limitations: Only certain domains were tested (e.g.: color-based, time, son and daughter, home and house, zero, impossibility metaphorical expressions, general Om and abu expressions, binomials, life and death, money, food and drink, Islamic expressions used metonymically, more discourse markers, derived forms from acronyms, clipping in daily speech, innovative word formation in Arabic, non-traditional plural forms on social media), deciphering unconventional Arabic spelling on social media; AI translations of full texts in other domain as). Only few AI models were included, mainly MC, DS, GT in most of the studies, Gemini in two studies, Ernie and Monica in one study each. Some domains had small sample sizes (e.g. 50 expressions). There is a need for multimodal datasets and culturally grounded corpora. Results reflect the models' performance at the time of testing (earliest in beginning of 2024 as medical terminology translation by AI). Replicating the studies now or later might render different results. Additionally, there is a need for longitudinal evaluation and for human-AI collaborative workflows.

4.11 *Future Research Directions*

Datasets in the studies reviewed herein suggest that future studies focus of the following: testing newer AI models as they evolve, expanding to proverbs and sayings in major Arabic dialects, studying AI performance on narrative translation, investigating AI's handling of irony, humor, and figurative language, comparing professional translators vs. AI, examining AI's ability to detect uncertainty or saying "I don't know",

4.12 *Positioning This Work Within Global AI Ethics and Bias Debates*

The findings of this review contribute directly to global debates on AI ethics, fairness, and linguistic bias. The systematic patterns observed across the twenty studies—particularly the dominance of English-centric training data, the underrepresentation of Arabic idioms and cultural expressions, and the models' reliance on literal translation defaults—highlight structural inequities in current AI development. These results demonstrate that AI systems do not fail randomly; they fail in ways that reflect global power imbalances in data availability, linguistic visibility, and cultural representation. By documenting how Arabic, a major world language, is systematically disadvantaged in AI translation performance, this study provides empirical evidence for the broader ethical argument that AI technologies reproduce and amplify existing linguistic hierarchies. Future research can build on this work by examining how culturally grounded datasets, community-driven annotation, and more equitable data distribution can mitigate these biases. In this way, the study positions Arabic linguistics as a critical site for understanding and addressing global AI bias.

5. Recommendations

Based on the SR and MA of the 20 AI studies conducted in the current study, this study offers some recommendations for translation pedagogy as integrating training modules that focus on culturally embedded expressions, metaphorical structures, and

metonymic names—domains where AI models consistently underperform; teaching translation-students strategies for dealing with uncertainty, including inferential reasoning and context-based interpretation, to reduce blank responses; incorporating comparative exercises where students analyze differences between human and AI translations to develop critical awareness, and strengthen instruction in scientific and medical terminology, where AI performs well, to help students build domain-specific competence.

Students should be trained to use inferential strategies rather than relying exclusively on dictionary lookups, and should be Encouraged to analyze AI errors as part of their translation training to understand linguistic weaknesses. They should use AI as a supplementary tool in scientific or standardized domains, but not as a primary resource for culturally rich or metaphorical content. Instructors should raise students' awareness that AI does not "understand" culture; it processes statistical patterns.

Students and the public should be taught how to identify linguistic and factual hallucinations in AI-generated texts. They should verify AI translations in politically sensitive, medical, or religious contexts before sharing them. Media literacy modules that explain the risks of relying on AI for interpreting culturally specific or historical content can be integrated.

Additionally it recommends the expansion of bilingual dictionaries to include Abu/Umm folk names, metonymic expressions, zero-expressions, and metaphorical grammatical terms that are currently under-documented; Providing culturally grounded definitions and usage notes for expressions that do not have direct equivalents in English; Including real contextual examples from spoken and written Arabic to support both human translators and AI systems and Developing lexicographic resources that reflect popular, colloquial, and folk usage—not only formal or scientific registers.

Thirdly, Arabic NLP research should enrich Arabic corpora with culturally loaded expressions, folk terminology, political terminology, and encrypted or colloquial varieties; develop models capable of handling metonymy, metaphor, and culturally anchored referents rather than relying solely on surface patterns; improve hallucination-detection mechanisms, especially in low-resource or culturally specific domains; and build more balanced bilingual datasets that reduce the dominance of English-centric training data.

Moderation teams should be trained to recognize that AI frequently misinterprets political terminology, coded expressions, and folk names and metaphors in Arabic. Automated tools can be developed to detect misleading AI translations circulating in encrypted platforms such as WhatsApp and Telegram. Specialized databases of culturally and politically sensitive terms can be built to reduce misclassification and mistranslation. Collaboration between NLP researchers and digital security experts should be encouraged to understand error patterns in encrypted Arabic communication.

6. Conclusion

This study offers the first comprehensive synthesis of AI performance in Arabic linguistics and translation across twenty empirical domains. By bringing together findings on scientific terminology, metaphorical expressions, folk expressions, political terminology, and calligraphic text recognition, the review provides a reference anchor for future scholarship in Arabic NLP, translation studies, and applied linguistics. The results highlight the role of AI as a promising tool and a persistent challenge: AI models excel in standardized, literal, and high-frequency domains, yet they consistently struggle with culturally embedded, metonymic, and context-dependent expressions that require deep linguistic grounding.

The synthesis demonstrates that AI's strengths lie in breadth of exposure, pattern recognition, and access to large multilingual corpora, while its weaknesses stem from limited cultural knowledge, shallow semantic processing, and hallucination under uncertainty. These findings establish a clear baseline for evaluating future models and underscore the need for culturally enriched datasets, improved morphological modelling, and more robust mechanisms for handling ambiguity.

As the first review study of AI's role in Arabic linguistics and translation, this work invites scholars to use it as a reference point for future studies, whether in translation pedagogy, dictionary building, NLP development, or media literacy research. By mapping what AI can and cannot do in Arabic, the study lays the foundation for more informed, culturally aware, and methodologically rigorous research in the future.

Conflicts of Interest: The author declares no conflict of interest.
ORCID ID: https://orcid.org/0000-0002-6255-1305

References

[1] Ahmad, M. M., Khan, I. A., & Tiwari, H. C. (2024). The Pros and Cons of Systematic Review and Meta-Analysis: Methodological Insights and Future Directions. *Digital Journal of Clinical Medicine*, *6*(4), 5.AlGhamedi, N. (2024). Constraints to neural machine translation quality, human and automated evaluation, and quality improvement across language pairs: A systematic literature review. *Journal of Research in Language & Translation (Special Issue), doi*, *10*.

[2] Ali, I., Warraich, N. & Butt, K. (2025). Acceptance and use of artificial intelligence and AI-based applications in education: A meta-analysis and future direction. *Information Development*, *41*(3), 859-874.

[3] Al-Jarf, R. (2025a). AI translation of full-text Arabic research articles: The case of educational polysemes. *Journal of Computer Science and Technology Studies*, 7(1), 311-325. Google Scholar

[4] Al-Jarf, R. (2025b). AI translation of the Gaza-Israel war terminology. *International Journal of Linguistics, Literature and Translation*, 8(2), 139-152. Google Scholar

[5] Al-Jarf, R. (2025c). Arabic transliteration of borrowed English nouns with /g/ by Artificial Intelligence (AI). *Journal of Computer Science and Technology Studies*, 7(9), 245-252. Google Scholar

[6] Al-Jarf, R. (2025d). Calligraphic Text Recognition by Gemini, Ernie ViLG and Google Translate: A Comparative Study of Arabic, Japanese and Chinese. *Journal of Computer Science and Technology Studies*, *7*(12), 474-494. https://doi.org/10.32996/jcsts.2025.7.12.54

[7] Al-Jarf, R. (2025e). Can AI decode and interpret encrypted Arabic on Facebook and YouTube to evade algorithmic moderation. *Journal of Computer Science and Technology Studies*, *7*(12), 307-321. DOI: 10.32996/jcsts.2025.7.12.40. Google Scholar

[8] Al-Jarf, R. (2025f). Can Artificial Intelligence (AI) translate Arabic abu-brand names with different prompts. *Journal of Computer Science and Technology Studies*, 7(9), 768-779. Google Scholar

[9] Al-Jarf, R. (2025i). Copilot's English translation of contrastive emphatic negation in Arabic discourse: An analytical study. *International Journal of Linguistics, Literature and Translation*, *8*(12), 214-230. DOI: 10.32996/ijllt.2025.8.12.24. Google Scholar

[10] Al-Jarf, R. (2025j). Copilot vs DeepSeek's translation of denotative and metonymic abu- and umm- animal and plant folk names in Arabic. *Journal of Computer Science and Technology Studies*, 7(10), 367-385. Google Scholar

[11] Al-Jarf, R. (2025k). DeepSeek, Google translate and Copilot's translation of Arabic grammatical terms used metaphorically. *Journal of Computer Science and Technology Studies*, 7(3), 46-57. Google Scholar

[12] Al-Jarf, R. (2025l). Google Translate then and now: Translations from five languages into English and Arabic (2012–2025). *Journal of Computer Science and Technology Studies*, *7*(12), 413-427. DOI: 10.32996/jcsts.2025.7.12.50. Google Scholar

[13] Al-Jarf, R. (2025m). Human vs AI translation of common names of chemical compounds: A comparative study. *Frontiers in Computer Science and Artificial Intelligence, 4*(4), 11-24. https://doi.org/10.32996/fcsai.2025.4.4.2. Google Scholar

[14] Al-Jarf, R. (2025o). Pronunciation errors in Arabic YouTube videos narrated by AI. *Frontiers in Computer Science and Artificial Intelligence*, *4*(2), 01-12. https://doi.org/10.32996/fcsai.2025.2.2.1. Google Scholar

[15] Al-Jarf, R. (2025q). Specific linguistic questions that Artificial Intelligence (AI) cannot answer accurately: Implications for Digital Didactics. *Frontiers in Computer Science and Artificial Intelligence*, *4*(4), 43-61. https://doi.org/10.32996/fcsai.2025.4.4.4. Google Scholar

[16] Al-Jarf, R. (2025r). *To publish or not to publish AI-generated research articles in scholarly journals: A perspective from editors and publishers*. I2COMSAPP International Conference on Artificial Intelligence and its Practical Applications in the Age of Digital Transformation. 2nd Edition. Faculty of Sciences and Techniques. Nouakchott University, Nouakchott, Mauritania. October 22-24, 2025. Google Scholar

[17] Al-Jarf, R. (2025s). Translation of Arabic expressions of impossibility by AI and student-translators: A comparative study. *Journal of Computer Science and Technology Studies*, 7(8), 288-299. Google Scholar

[18] Al-Jarf, R. (2025t). Translation of Arabic folk medical terms with om and abu by AI: A comparison of Microsoft Copilot and DeepSeek. *Journal of Medical and Health Studies, 6*(4), 45-58. Google Scholar

[19] Al-Jarf, R. (2025u). Translation of English and Arabic "sleep" terms and formulaic expressions by Artificial Intelligence: A comparison of Copilot and DeepSeek. *International Journal of Linguistics, Literature and* Translation, 8(11), 95-108. Google Scholar

[20] Al-Jarf, R. (2025v). Translation of zero-expressions by Microsoft Copilot and Google Translate. *Journal of Computer Science and Technology Studies*, 7(2), 203-216. Google Scholar

[21] Al-Jarf, R. (2025w). Faulty consonant gemination in the pronunciation of English biomedical terms by Arab healthcare professionals. *Journal of Medical and Health Studies, 6*(3), 56-66. https://doi.org/10.32996/jmhs.2025.6.3.9. Google Scholar

[22] Al-Jarf, R. (2025x). Splitting unsplittable foreign words in casual speech by EFL Arab Learners. *British Journal of Applied Linguistics, 5*(2), 01-11. https://doi.org/10.32996/bjal.2025.5.2.1 Google Scholar

[23] Al-Jarf, R. (2025y). Vowel pronunciation errors in English biomedical terminology by Arab healthcare professionals. *Journal of Medical and Health Studies, 6*(2), 145-155. https://doi.org/10.32996/jmhs.2025.6.2.22. Google Scholar

[24] Al-Jarf, R. (2025z). Mapping pronunciation errors in English silent consonants: A Corpus-based Study of Saudi EFL Undergraduates. *Journal of Humanities and Social Sciences Studies, 7*(6), 13-21. DOI: 10.32996/jhsss.2025.7.6.2. Google Scholar

[25] Al-Jarf, R. (2024a). Expressions of impossibility in Arabic and English: Unveiling students' translation difficulties. *International Journal of Linguistics, Literature and Translation*, *7*(5), 68-76. DOI: 10.32996/ijllt.2024.7.5.9. ERIC ED651472. Google Scholar

[26] Al-Jarf, R. (2024b). Students' assignments and research papers generated by AI: Arab instructors' views. *Journal of Computer Science and Technology Studies*, 6(2), 92-98. Google Scholar

[27] Al-Jarf, R. (2024c). The Gaza-Israel war terminology: implications for translation pedagogy. *International Journal of Middle Eastern Research, 3*(1), 35-43. DOI: 10.32996/ijmer.2024.3.1.5. ERIC ED650283 Google Scholar

[28] Al-Jarf, R. (2024d). To translate or not to translate: The case of Arabic and foreign shop names in Saudi Arabia. *International Journal of Translation and Interpretation Studies, 4*(1), 33-40. DOI: 10.32996/ijtis.2024.4.1.5. Google Scholar

[29] Al-Jarf, R. (2024e). *Translation of medical terms by AI: A comparative linguistic study of Microsoft Copilot and Google Translate*. In Y. M. Elhadj et al. (Eds.), I2COMSAPP 2024, LNNS 862, pp. 1–16. https://doi.org/10.1007/978-3-031-71429-0_17. Springer Nature Switzerland AG 2024. Google Scholar

[30] Al-Jarf, R. (2023a). Clipping of borrowings in spoken Arabic. International Journal of Linguistics, Literature and Translation, 6(1), 68-76. ERIC ED633842 https://doi.org/10.32996/ijllt.2023.6.11.9

[31] Al-Jarf, R. (2023b). Word formation with foreign lexemes: The case of hybrid compounds in Arabic. *Journal of Humanities and Social Sciences Studies, 5*(11), 15–27. https://doi.org/10.32996/jhsss.2023.5.11.3.

[32] Al-Jarf, R. (2023c). Equivalence problems in translating ibn (son) and bint (daughter) fixed expressions to Arabic and English. *International Journal of Translation and Interpretation Studies, 3, 2, 1-15*. DOI: 10.32996/ijtis.2023.3.2.1. ERIC ED628181 Google Scholar

[33] Al-Jarf, R. (2023d). Numeral-based English and Arabic formulaic expressions: cultural, linguistic and translation issues. *British Journal of Applied Linguistics, 3,1,* 25-34. https://doi.org/10.32996/bjal.2023.3.1.2. ERIC ED628151. Google Scholar

[34] Al-Jarf, R. (2023e). Semantic and syntactic anomalies of Arabic-transliterated compound shop names in Saudi Arabia. *International Journal of Arts and Humanities Studies* (IJAHS), 3(1), 1-8. DOI: 10.32996/ljahs.2023.3.1.1. Google Scholar

[35] Al-Jarf, R. (2023f). Time metaphors in English and Arabic: Translation challenges. *International Journal of Translation and Interpretation Studies (IJTIS), 3*(4), 68-81 https://doi.org/10.32996/ijtis.2023.3.4.8. Google Scholar

[36] Al-Jarf, R. (2022a). Arabic and English dar (house) and bayt (home) expressions: Linguistic, translation and cultural issues. *Journal of Pragmatics and Discourse Analysis (JPDA), 1*(1), 1-13. ERIC ED624367 Google Scholar

[37] Al-Jarf, R. (2022b). Challenges that undergraduate student translators' face in translating polysemes from English to Arabic and Arabic to English. *International Journal of Linguistics, Literature and Translation (IJLLT), 5(7*), 84-97. DOI: 10.32996/ijllt.2022.5.7.10. ERIC ED620804. Google Scholar

[38] Al-Jarf, R. (2022c). Deviant Arabic transliterations of foreign shop names in Saudi Arabia and decoding problems among shoppers. *International Journal of Asian and African Studies (IJAAS), 1*(1), 17-30. DOI: 10.32996/ijaas.2022.1.1.3. Google Scholar

[39] Al-Jarf, R. (2022d). English transliteration of Arabic personal names with the definite article /al/ on Facebook. *British Journal of Applied Linguistics (BJAL), 2*(2), 23-37. DOI: 10.31926/but.pcs.2022.64.15.2.2. Google Scholar

[40] Al-Jarf, R. (2022e). Gemination errors in Arabic-English transliteration of personal names on Facebook. *International Journal of Linguistics Studies (IJLS), 2*(2),163-170. DOI: 10.32996/ijls.2022.2.2.18. Google Scholar

[41] Al-Jarf, R. (2022f). Issues in Translating English and Arabic common names of chemical compounds by student-translators in Saudi Arabia. In Kate Isaeva (Ed.). Special Knowledge Mediation: Ontological & Metaphorical Modelling. Springer. DOI: 10.1007/978-3-030-95104-7. Google Scholar

[42] Al-Jarf, R. (2022g). Proper noun pronunciation inaccuracies in English by Educated Arabic speakers. *British Journal of Applied Linguistics (BJAL), 4*(1), 14-21. https://doi.org/10.32996/bjal.2022.2.1.3. **ERIC** ED619388. Google Scholar

[43] Al-Jarf, R. (2022h). Undergraduate student-translators' difficulties in translating English word + preposition collocations to Arabic. International Journal of Linguistics Studies (IJLS), 2(2), 60-75. DOI: 10.32996/ijls.2022.2.2.9. ERIC ED621368. Google Scholar

[44] Al-Jarf, R. (2022i). Text-to-speech software for promoting EFL freshman students' decoding skills and pronunciation accuracy. *Journal of Computer Science and Technology Studies (JCSTS), 4*(2), 19-30. DOI: 10.32996/jcsts.2022.4.2.4. ERIC ED621861. Google Scholar

[45] Al-Jarf, R. (2022j). Text-to-speech software as a resource for independent interpreting practice by undergraduate interpreting students. *International Journal of Translation and Interpretation Studies (IJTIS), 2*(2), 32-39. DOI: 10.32996/ijtis.2022.2.2.3. ERIC ED621859. Google Scholar

[46] Al-Jarf, R. (2022k). Variant transliterations of the same Arabic personal names on Facebook. *International Journal of English Language Studies (IJELS), 4*(4), 79-90. DOI: 10.32996/ijels.2022.4.4.11. Google Scholar

[47] Al-Jarf, R. (2022l). YouTube videos as a resource for self-regulated pronunciation practice in EFL distance learning environments. *Journal of English Language Teaching and Applied Linguistics (JELTAL), 4*(2), 44-52. https://doi.org/10.32996/jeltal.2022.4.2.4. ERIC ED618965. Google Scholar

[48] Al-Jarf, R. (2021a). An investigation of Google's English-Arabic translation of technical terms. *Eurasian Arabic Studies*, 14, 16-37. Google Scholar

[49] Al-Jarf, R. (2021b). Arabic and English loan words in Bahasa: Implications for Foreign Language Pedagogy. *Journal La Edusci, 2*(4), 23-35. https://doi.org/10.37899/journallaedusci.v2i4.44

[50] Al-Jarf, R. (2021c). Mobile audiobooks, listening comprehension and EFL college students. *International Journal of Research – GRANTHAALAYAH, April 9*(4), 410-423. https://doi.org/10.29121/granthaalayah.v9.i4.2021.3868. Google Scholar

[51] Al-Jarf, R. (2021d). TED Talks as a Listening Resource in EFL College classrooms. International Journal of Language and Literary Studies (IJLLS), 2(3), 256–267. https://doi.org/10.36892/ijlls.v2i3.691. ERIC ED615127.

[52] Al-Jarf, R. (2020). Issues in translating English and Arabic plurals. Universitatea „1 Decembrie 1918" din Alba Iulia - The *Journal of Linguistic and Intercultural Education - JoLIE, 13*(1), 7-28. https://doi.org/10.29302/jolie.2020.13.1. Google Scholar

[53] Al-Jarf, R. (2019a). EFL Freshman students' difficulties with phoneme-grapheme relationships. 5th VietTESOL International Convention. Hue University of Foreign Languages, Hue, Vietnam. October 11-12. Google Scholar

[54] Al-Jarf, R. (2019b). Translation students' difficulties with English and Arabic color-based metaphorical expressions. *Fachsprache, 41* (Sp. Issue), 101-118. Doi: 10.24989/fs.v41iS1.1774. ERIC ED622935. Google Scholar

[55] Al-Jarf, R. (2018a). Effect of Background Knowledge on Auditory Comprehension in Interpreting Courses. In Renata Jancarikova (Ed.) Interpretation of Meaning across Discourse, pp. 97-108. Muni Press, Brno, Czech Republic. https://doi.org/10.5817/CZ.MUNI.M210-8947-2018. ERIC ED665097. Google Scholar

[56] Al-Jarf, R. (2018b). Multiple Arabic equivalents to English medical terms: Translation issues. *International Linguistics Research, 1(*1); 102-110: 2018. https://doi.org/10.30560/ilr.v1n1p102. ERIC ED613073. Google Scholar

[57] Al-Jarf, R. (2017a). Issues in translating Arabic om- and abu-expressions. Alatoo Academic Studies, 3, 278-282. ERIC ED613247. Google Scholar

[58] Al-Jarf, R. (2017b). Technology Integration in Translator Training in Saudi Arabia. International Journal of Research in Engineering and Social Sciences (IJRESS), 7(3) (March), 1-7. ERIC ED613071. Google Scholar

[59] Al-Jarf, R. (2016a). *Issues in translating English technical terms to Arabic by Google Translate*. *TICET 2016 Conference*, Khartoum, Sudan. Google Scholar

[60] Al-Jarf (2016b). *Translation of English and Arabic binomials by advanced and novice student translators*. In Larisa Ilynska and Marina Platonova (Eds) Meaning in Translation: Illusion of Precision (Pp. 281-298). Cambridge Scholars Publishing. ERIC ED639264. Google Scholar

[61] Al-Jarf, R. (2012). *Electronic translation between Arabic and European languages: Current status and future Perspectives*. 6th Annual Conference of Ibn Sina Institute for Human Sciences titled: The Future of Arabic Language Teaching in Europe. LILLE, France. June 22-24. Google Scholar

[62] Al-Jarf, R. (2010a). *Interlingual pronoun errors in English-Arabic translation*. International symposium on Using Corpora in Contrastive and Translation Studies. Edge Hill University, UK. Google Scholar

[63] Al-Jarf, R. (2010b). Translation students' difficulties with English neologisms. *Analele Universităţii "Dunărea De Jos" Din Galaţi Fascicula XXIV ANUL II*I, 2, 431-437. Romania. Google Scholar

[64] Al-Jarf, R. (2009a). *How to use the OmegaT translation memory*. Sultan Qaboos University, Muscat, Oman. April 12-13. Google Scholar

[65] Al-Jarf, R. (2009b). *Word+particle collocation errors in English-Arabic translation*. 40 Years of Particle Research. Bern, Switzerland. February 11.-13. Google Scholar

[66] Al-Jarf, R. (2007). SVO word order errors in English-Arabic translation. *META, 52*(2), 299–308. DOI:10.7202/016072ar. ERIC ED623835. Google Scholar

[67] Al-Jarf, R. (2005a). The effects of listening comprehension and decoding skills on spelling achievement of EFL Freshman Students. English language and literature Education. *Journal of the English Language Teachers in Korea (ETAK), 11*(2). ERIC ED625524. Google Scholar

[68] Al-Jarf, R. (2005b). The *relationship among spelling, listening and decoding skills* in EFL *freshman students*. *English Language & Literature Teaching, 11(2), 35-55.* Google Scholar

[69] Al-Jarf, R. (2000). Grammatical agreement errors in L1/L2 translation. *International Review of Applied Linguistics, 38,* 1-15. https://doi.org/10.1515/iral.2000.38.1.1. Google Scholar

[70] Bangdiwala, S. I. (2024). The importance of systematic reviews. *International Journal of Injury Control and Safety Promotion*, *31*(3), 347–349. https://doi.org/10.1080/17457300.2024.2388484

[71] Batubara, M. et al. (2025). A Systematic Literature Review on the Application of Artificial Intelligence in Translation: Challenges, Innovations, and Impact Across Diverse Fields. *Journal of Linguistics, Literature, and Language Teaching (JLLLT)*, *5*(1), 67-87.

[72] Chan, V., & Tang, W. (2024). GPT for translation: A systematic literature review. *SN computer science*, *5*(8), 986.

[73] Chen, A. et al. (2025). A systematic review and meta-analysis of AI-enabled assessment in language learning: Design, implementation, and effectiveness. *Journal of Computer Assisted Learning*, *41*(1), e13064.
[74] Deng, X., & Yu, Z. (2022). A systematic review of machine-translation-assisted language learning for sustainable education. *Sustainability*, *14*(13), 7598.
[75] Elsadig, M. (2024). The Impact of Artificial Intelligence on Language Translation: A Review. IEEEAccess. DOI: 10.1109/ACCESS.2024.3366802
[76] Kim, G. (2023). How to perform and write a systematic review and meta-analysis. *Child Health Nursing Research*, *29*(3), 161.
[77] Lee, S. (2023). The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning*, *36*(1-2), 103-125.
[78] Mai, M., Mirza, F. & DiMarco, C. (2024). Application of text-to-image translation algorithms in medicine: a systematic review. *JAAD Reviews*, *2*, 88-96.
[79] McNaughton, J. et al. (2023). Machine learning for medical image translation: A systematic review. *Bioengineering*, *10*(9), 1078.
[80] Mirzaeian, V. & Oskoui, K. (2023). Google Translate in foreign language learning: A systematic review. *Applied Research on English Language*, *12*(2), 51-84.
[81] Mohamed, Y. et al. (2024). The impact of artificial intelligence on language translation: a review. *Ieee Access*, *12*, 25553-25579.
[82] Nguyen, T. et al. (2025). The Benefits and Challenges of AI Translation Tools in Translation Education at the Tertiary Level: A Systematic Review. *International Journal of TESOL & Education*, *5*(2), 132-148.
[83] Noll, R. et al. (2023). Machine translation of standardised medical terminology using natural language processing: a scoping review. *New biotechnology*, *77*, 120-129.
[84] Omar, L. & Salih, A. (2024). Systematic review of English/Arabic machine translation postediting: Implications for AI application in translation research and pedagogy. In *Informatics* (Vol. 11, No. 2, p. 23). MDPI.
[85] Ssemugabi, S. (2025). The role of AI in Modern Language Translation and its Societal Applications: A Systematic literature review. In *Southern African Conference for Artificial Intelligence Research* (pp. 390-404). Springer, Cham.
[86] Sun, A., & Han, W. (2025). A Systematic Review of Financial Translation in the Digital Age: Trends, Challenges, and Human–AI Collaboration.
[87] Tafa, T. et al. (2025). Machine translation Performance for LowResource Languages: A Systematic Literature Review. *IEEE Access*.
[88] Teibowei, M. & Mbete, T. (2023). Meta-Analysis of the Efficacy of Machine translation in Biomedical Texts. International Journal of Medical Evaluation and Physical Report, Vol 7. No. 4 www.iiardjournals.org
[89] Yang, J. (2025). A systematic review of online AI translation tools for English language learning and teaching. *International Journal of Internet, Broadcasting and Communication*, 129-139.
[90] Yazar, B., Şahın, D. & Kiliç, E. (2023). Low-resource neural machine translation: A systematic literature review. *IEEE Access*, *11*, 131775-131813.
[91] Zappatore, M., & Ruggieri, G. (2024). Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, *84*, 101582.
[92] Yuan Y, Hunt RH. (2009). Systematic reviews: the good, the bad, and the ugly. Am J Gastroenterol, 104(5):1086-92. doi: 10.1038/ajg.2009.118. PMID: 19417748.