Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Conversational Lexicons in Human–AI Interaction: A Mixed-Methods Study of Generative AI–Supported Lexical Decision-Making

Mohamed Mekheimer

mohamed.abdelgawad3@edu.bsu.edu.eg

Beni-Suef University

Fatma Abdelaal

Beni-Suef University

Research Article

Keywords: generative AI, human–AI interaction, natural-language interfaces, translation technology, cognitive load, AI literacy

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11255152/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Conversational Lexicons in Human–AI Interaction: A Mixed-Methods Study of Generative AI–Supported Lexical Decision-Making

Mohamed Mekheimer, PhD (Corresponding Author)
Professor of TESOL
Faculty of Education, Beni-Suef University, Beni Suef City, Egypt
Email: mohamed.abdelgawad3@edu.bsu.edu.eg
ORCID: https://orcid.org/0000-0001-6784-0993

Fatma Abdelaal, PhD
Associate Professor of Teaching French as a Foreign Language
Faculty of Education, Beni-Suef University, Beni Suef City, Egypt
Email: fatma.abdelaal1@edu.bsu.edu.eg
ORCID: https://orcid.org/0009-0008-5361-052X

Conversational Lexicons in Human–AI Interaction: A Mixed-Methods Study of Generative AI–Supported Lexical Decision-Making

Abstract

Large language models are increasingly used as interactive language tools, yet their effects on fine-grained human–AI decision-making remain underexamined. This mixed-methods experiment investigated a ‘conversational lexicon’ workflow in which student translators used a generative AI system to explore lexical alternatives while retaining responsibility for final decisions. Sixty senior Arabic-speaking student translators working from English or French into Arabic were randomized within language cohort to a GenAI-supported condition (n = 30) or a traditional digital-resource condition (n = 30). Outcomes included final translation quality, perceived cognitive load, creative confidence, pre-task self-reported GenAI-use functions, screen/audio records, and interviews. The GenAI group obtained higher quality scores than the control group (M = 46.47 vs. 35.40), with Welch’s $t(44.62) = 51.19$, $p < .001$. It also reported lower cognitive load and higher creative confidence; a joint multivariate test was significant, $F(2, 57) = 554.24$, $p < .001$. Pre-task responses showed that contextual explanation and alternative phrasing were the most frequently reported GenAI uses. Qualitative evidence highlighted iterative prompting, cross-checking, contextual judgment, and concerns about over-reliance. The findings support a human-centered interaction account in which natural-language interfaces expand the option space for lexical decisions but require active verification. The unusually large between-group effects should be interpreted cautiously because outcomes showed restricted within-group dispersion and arose from a tightly controlled, task-specific intervention.

Keywords: generative AI; human–AI interaction; natural-language interfaces; translation technology; cognitive load; AI literacy

Introduction

Generative artificial intelligence (GenAI) is reshaping translation practice through natural-language interfaces that allow users to query, refine, challenge, and contextualize model outputs. From a computing perspective, this is a human–AI interaction problem: the relevant unit is not only model output quality, but also how users formulate requests, evaluate alternatives, calibrate trust, and incorporate generated suggestions into task workflows. Translation provides a demanding setting for studying these interaction dynamics because lexical decisions require simultaneous attention to meaning, register, context, and cultural fit.

This study focuses on the micro-level interaction through which users solve lexical problems. Traditional digital dictionaries mainly support retrieval from relatively fixed lexical resources, whereas large language models (LLMs) support iterative, context-sensitive generation. We use the term ‘conversational lexicon’ to describe this interface function: an on-demand natural-language resource through which a user can request explanations, compare candidate expressions, specify register or tone, and iteratively refine a lexical solution while retaining final decision authority.

The distinction matters because conversational querying changes the interaction loop. A request such as ‘What is a more poetic way to say X?’ does not merely retrieve a stored equivalent; it elicits generated alternatives conditioned on contextual instructions. Such interaction may reduce search fragmentation and support creative exploration, but fluent outputs can also be misleading, culturally mismatched, or overly generic. The resulting computing challenge is therefore one of human-centered augmentation: determining whether and how an LLM interface changes user performance, cognitive effort, confidence, verification behavior, and trust calibration.

These questions are especially salient in translation into Arabic from English and French, where lexical choices often require sensitivity to morphology, diglossia, register, idiomaticity, and culturally specific meaning. The present study therefore uses translation as an empirical testbed for human–AI lexical decision-making. It addresses three questions:

1. 1. What lexical functions do student translators report using GenAI for before the experimental task?
2. 2. Under the experimental conditions, how is access to a conversational-lexicon workflow associated with translation quality, perceived cognitive load, and creative confidence?
3. 3. How do participants describe the benefits, risks, and verification strategies involved in conversational lexical interaction with GenAI?

Literature Review

GenAI has rapidly become a central concern in Translation Studies, with recent work examining its implications for professional workflows, translator training, competence development, EFL learning, and industry productivity [13, 26, 31, 39, 42]. Existing studies document new forms of human-machine collaboration [17, 25, 31, 35], evaluate integration into CAT-tool ecosystems, and compare GenAI-assisted translation with conventional machine translation in terms of speed and quality [4, 28, 33]. What remains less fully examined is the smaller but pivotal interaction through which translators negotiate word choice itself.

From Static Look-up to Dynamic Inquiry: The Evolving Lexicographical Landscape

The traditional dictionary has long been understood as a finite resource for lexical correspondence [32]. GenAI challenges this model by enabling a dynamic, context-sensitive form of lexical engagement. Lew [27] identifies this transformation as a major issue for contemporary lexicography, while Cai and Tian [14] show that translators' GenAI-based information seeking differs markedly from conventional web or dictionary searches.

Rather than asking for a single equivalent, translators can now use conversational prompting to explore a range of lexical possibilities [23, 40]. Such prompts may request stylistic alternatives, register-specific terminology, culturally attuned phrasing, or explanations of connotation. The interaction consequently shifts from retrieval to exploration: the translator uses the tool to test possibilities and refine judgment rather than merely to locate a predetermined answer.

This shift is enabled by the ability of LLMs to process context-rich queries. Unlike translation memories or termbases that depend on exact or fuzzy matching, LLMs can draw on surrounding linguistic and situational information to generate tailored suggestions [22, 34]. Their capacity for contextual disambiguation [12] and dynamic semantic generation [20, 44] allows them to offer contextually plausible definitions, collocations, and usage examples [29].

As a result, the translator's relationship with lexical tools is being redefined. The LLM becomes less a passive repository than an interactive linguistic partner embedded in the cognitive workflow [26, 37]. The process becomes a collaborative exploration of language in which the translator uses AI-generated options to extend, question, and refine professional judgment [21].

The Human-AI Symbiosis: Cognitive and Creative Implications

This emergent human-AI symbiosis positions the translator as the curator of a creative dialogue rather than the operator of a productivity tool. By allowing nuanced, context-specific prompts, GenAI can support creative exploration, help resolve translator's block, and broaden access to stylistic alternatives that traditional thesauri rarely provide [21, 23, 40]. In this sense, the conversational lexicon encourages deeper lexical engagement by generating options that the translator must compare, adapt, and evaluate [34, 46].

At the same time, the same fluency that makes GenAI useful also introduces professional and cognitive risks. Over-reliance on statistically probable outputs may produce "lexical smoothing," reducing stylistic distinctiveness and encouraging generic phrasing [21]. AI-generated suggestions may also be factually inaccurate, culturally tone-deaf, or idiomatically unnatural, particularly in contexts that demand careful attention to cultural nuance [3, 12]. These risks shift part of the translator's work toward evaluation, verification, and risk assessment [31, 40, 42, 45].

For this reason, successful integration depends on AI literacy: the ability not only to prompt effectively but also to recognize when an AI suggestion should be accepted, revised, or rejected [26, 29, 45]. The conversational lexicon therefore enhances the

translator's resources while simultaneously increasing the importance of critical judgment.

The Gap: Cross-Linguistic Nuance and the Need for Focused Investigation

Although the literature establishes that GenAI is changing translation practice, much of it treats translation in linguistically general terms. This leaves underexplored how the conversational lexicon functions in specific language pairs, where structural distance, cultural reference, stylistic convention, and register shape both the usefulness and the risks of AI-generated suggestions [15, 45].

Translation into Arabic illustrates these complexities. In English-Arabic translation, lexical decisions are shaped by Arabic morphology, diglossia, and a rich inventory of culture-specific and religious idioms for which direct equivalents are often unavailable [3, 12]. In this context, an LLM may help generate contextual options and conceptual bridges, but it may also produce expressions that are plausible yet culturally inappropriate or tonally inaccurate [7, 36].

French-Arabic translation raises a related but distinct set of issues. Because French translation often places a premium on register, nuance, and le mot juste, a conversational lexicon may be used to explore fine-grained stylistic alternatives rather than to locate missing cultural equivalents [23]. This creates a productive tension between GenAI's creative flexibility and the need to maintain formal conventions, terminological consistency, and appropriate register [21, 25].

Considered together, these studies demonstrate the need for a focused empirical investigation of GenAI-mediated lexical decision-making in translation. Recent research confirms that AI is increasingly reshaping translation practice by accelerating lexical search, expanding access to alternative renderings, and supporting multilingual text production [5, 30]. However, this transformation is not uniform across language pairs, genres, or translation problems. Arabic–English and Arabic–French translation remain especially sensitive to ambiguity, idiomaticity, code-switching, legal precision, cultural nuance, and domain-specific terminology [8, 9, 10, 16]. These challenges suggest that translators working from English or French into Arabic may use similar GenAI functions—such as asking for equivalents, paraphrases, explanations, usage examples,

or contextual distinctions—but the cognitive demands and verification procedures associated with those functions may vary according to the source language and the lexical problem being solved.

The need for such investigation is further reinforced by evidence that AI systems remain unreliable in answering certain linguistic questions, particularly those requiring fine-grained pragmatic, semantic, cultural, or context-sensitive judgment [6]. Studies of AI-assisted translation also show that human expertise remains essential for resolving ambiguity, evaluating stylistic appropriateness, and maintaining ethical and professional responsibility in specialized translation contexts [2, 19]. At the same time, translation and interpreting research has long emphasized that lexical decisions are not merely mechanical substitutions but cognitively demanding acts involving attention, memory, inference, reformulation, and strategy selection [1]. Emerging work on AI-assisted translation and cognitive load similarly indicates that digital support may reduce some processing burdens while introducing new demands related to verification, trust calibration, and decision monitoring [38].

In summary, existing research establishes that GenAI is transforming translation, but it has not yet sufficiently examined the emergent practice of using large language models as dynamic conversational lexical tools within complex cross-linguistic workflows. The present study addresses this gap by analyzing student translators' conversational queries, the extent to which these queries shape cognitive load and creative confidence, and the verification strategies students develop when using GenAI for lexical decision-making across English–Arabic and French–Arabic translation tasks. This focus moves beyond general claims about AI-assisted translation by examining how student translators actually interact with GenAI at the level of lexical problem-solving, where linguistic ambiguity, source-language variation, and human evaluative judgment converge.

Theoretical Framework

This study is organized by the Conversational Lexicon Mediation Model (CLMM) (Fig. 1), a human–AI interaction framework for lexical decision-making. The model treats GenAI neither as a static dictionary nor as an autonomous translator, but as an interactive computational resource embedded in a user-controlled decision loop. It

integrates tool-mediated cognition, human–AI collaboration, cognitive load, translation competence, and AI literacy to explain how natural-language interaction can alter both task performance and user experience.

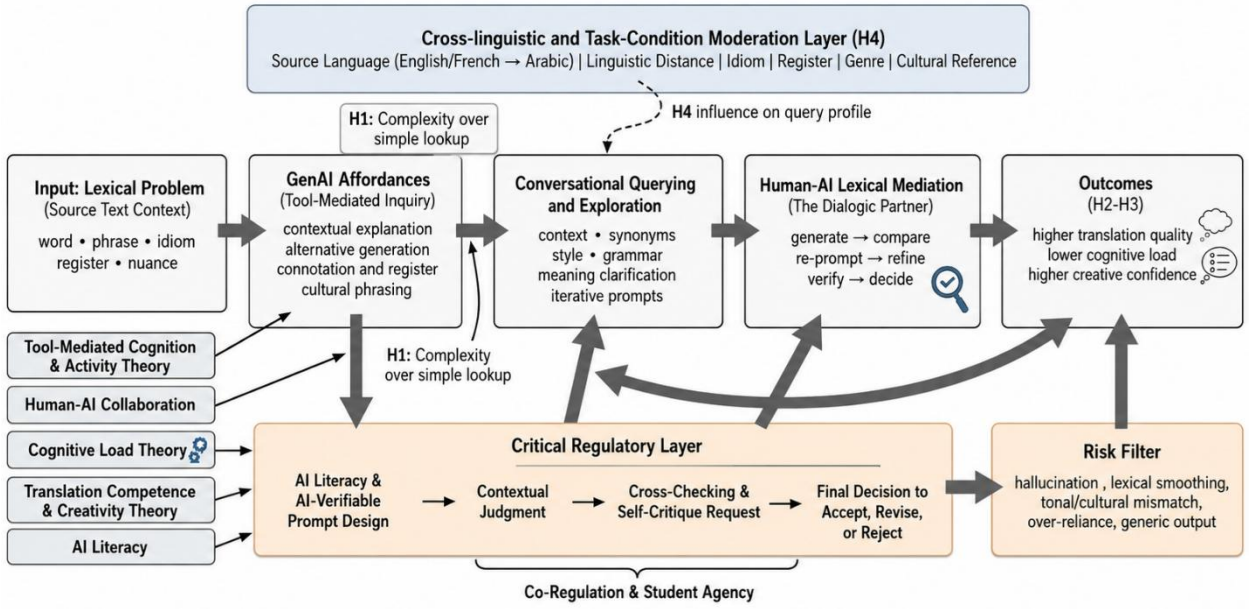


Fig. 1 The Conversational Lexicon Mediation Model (CLMM): A Framework for AI-Reshaped Lexical Decisions

The first theoretical assumption is that lexical inquiry is tool-mediated. From the perspective of tool-mediated cognition and activity theory, translators do not solve lexical problems in isolation; they act through symbolic and technological tools that shape what they notice, how they search, and what options become available. A traditional dictionary supports retrieval from a relatively fixed lexical database, whereas a conversational LLM supports exploratory interaction. This distinction is central to the present study because the literature review shows that GenAI changes the translator's information-seeking behavior from static lookup to dynamic inquiry, allowing prompts for contextual explanation, register, connotation, stylistic alternatives, and culturally attuned phrasing.

The second assumption is that GenAI creates a human-AI lexical mediation process. In this process, the translator begins with a lexical problem arising from the source text, formulates a prompt, receives a set of AI-generated options, and then compares, adapts, verifies, or rejects those options. The LLM is therefore conceptualized as a conversational lexicon: a dialogic partner embedded in the translator's cognitive workflow. The translator's agency remains central

because the AI expands the lexical option space, but the translator remains responsible for contextual appropriateness, cultural fit, and stylistic coherence.

The third assumption concerns cognitive and creative mediation. Cognitive load theory helps explain why a conversational lexicon may reduce perceived mental effort: instead of fragmenting attention across multiple dictionaries, search engines, thesauri, and context tools, the translator can conduct several forms of lexical exploration within one interactional space. This may reduce extraneous load and allow more attention to be directed toward higher-order stylistic and semantic decisions. At the same time, creativity theory explains why AI-generated alternatives may increase creative confidence. By producing multiple candidate phrasings, metaphors, collocations, and register-sensitive options, GenAI can function as a generative stimulus that helps students overcome lexical blocks and test possibilities they might not have produced independently.

The fourth assumption is regulatory. The benefits of GenAI are not automatic; they depend on AI literacy and verification. Because LLM outputs can be fluent but inaccurate, generic, culturally tone-deaf, or stylistically unsuitable, the model includes a critical regulatory layer. This layer consists of prompt design, cross-checking, iterative re-prompting, self-critique requests, contextual judgment, and the final decision to accept, revise, or reject a suggestion. This layer protects against hallucination, lexical smoothing, over-reliance, and tonal or cultural mismatch. Thus, the model does not equate GenAI use with outsourcing translation judgment; rather, it defines effective use as disciplined human-AI co-regulation.

A final boundary condition is that language pair, genre, register, and cultural reference may influence the usefulness and risk of generated lexical options. These contextual factors are retained as theoretically important moderators, but they are not tested as a separate confirmatory hypothesis here because the analysis archive does not contain a row-level categorical source-language variable.

In combination, the CLMM motivates three testable expectations. First, if users treat GenAI as more than a translation engine, their reported uses should emphasize contextual and generative lexical functions. Second, access to the conversational interface should be associated with higher evaluated task performance under the study conditions. Third, consolidating multiple forms of lexical exploration within one interface

should be associated with lower perceived cognitive load and higher creative confidence. Verification behavior is examined qualitatively as the regulatory mechanism that keeps the interaction human-centered.

Hypotheses

Based on the literature, the following hypotheses were formulated:

H1: Before the experimental task, participants will report using GenAI more frequently for context-sensitive and generative lexical functions (e.g., understanding context and finding alternative phrasing) than for full-sentence translation.

H2: Participants assigned to the GenAI-supported conversational-lexicon condition will obtain higher final translation-quality scores than participants assigned to traditional digital lexical resources.

H3: Participants assigned to the GenAI-supported condition will report lower perceived cognitive load and higher creative confidence than participants assigned to traditional digital lexical resources.

Methods and Materials

This study used a randomized, explanatory mixed-methods design to examine a human–AI lexical interaction workflow. Within each source-language cohort, participants were assigned to either a GenAI-supported conversational-lexicon condition or a traditional digital-resource condition. Quantitative analyses compared task products and post-task perceptions, while screen/audio records and interviews were used to interpret interaction strategies, verification practices, and perceived risks. The pre-task GenAI-use checklist was analyzed separately as a background usage profile and was not treated as observed behavior during the intervention.

Participants

A total of sixty senior undergraduate student translators participated in the study. The sample was composed of two distinct cohorts: thirty students from an English Department translating from English into Arabic, and thirty students from a French Department translating from French into Arabic. Inclusion criteria required participants to be native speakers of Arabic

enrolled in a final-year translation practicum course, ensuring they possessed a foundational level of translation competence.

All participants reported high proficiency in their respective source languages and prior training with computer-assisted translation (CAT) tools. Recruitment was conducted through departmental announcements. Within each language cohort, 15 participants were allocated to the control condition and 15 to the GenAI-supported condition, yielding n = 30 per experimental condition.

The study was conducted in accordance with the Declaration of Helsinki and approved by the Faculty of Education Institutional Review Board (FoE Ethics Committee), Beni-Suef University, Beni Suef City, Egypt (Approval No. FoE-002-02-25). All participants provided written informed consent before their inclusion in the study.

Materials and Instruments

The study used source texts, digital tools, questionnaires, process recordings, expert product scoring, and a semi-structured interview protocol. The full questionnaires and interview protocol are provided in Supplementary Information 1, and the de-identified analysis dataset and legacy statistical-output files are supplied as Supplementary Data files.

1. **Source Texts:** Two literary excerpts of approximately 300 words each were selected for their stylistic richness and nuanced vocabulary, ensuring the translation task would necessitate lexical research.
    - The English source text was excerpted from a contemporary literary novel.
    - The French source text was excerpted from a work of comparable difficulty and lexical complexity.
    - **Validation:** To ensure task suitability and equivalence, both texts were pre-validated by two independent bilingual experts for their respective language pairs (English-Arabic and French-Arabic).

2. **Technology and Tools:**
    - **Control Group:** Participants worked within a standard web-based CAT tool and were provided with a curated list of high-quality online resources, including bilingual dictionaries (e.g., WordReference), context dictionaries (e.g., Reverso Context), and online thesauri. They were explicitly prohibited from using GenAI tools.
    - Experimental Group: Participants used the same CAT environment and had access to ChatGPT-4 in a separate browser window. The system was framed as an interactive lexical resource rather than a full-text translation service, allowing users to request explanations, alternatives, and contextual distinctions.
3. **Data Collection Instruments:**
    - **Screen and Audio Recording Software:** All on-screen activity and audio from the think-aloud protocol were captured to analyze the translation process in real-time.
    - Pre- and Post-Task Questionnaires: The pre-task instrument collected demographic, language-background, technology-use, and self-reported GenAI-use information. The post-task instrument contained four cognitive-load items and four creative-confidence items rated on five-point scales. Recalculation from the uploaded raw dataset yielded Cronbach's $\alpha = .926$ for cognitive load and $\alpha = .929$ for creative confidence. The instruments are reproduced in Supplementary Information 1.
    - Expert Scoring of Translations: Final translations were anonymized and evaluated by two independent expert raters using a 50-point rubric covering accuracy, fluency, and stylistic appropriateness. The analysis archive supplied for this revision contains only the final aggregate scores, not rater-level scores; therefore, inter-rater reliability was not independently recalculated for the present manuscript and no unverifiable reliability coefficient is reported.

- Semi-Structured Interview Protocol: Follow-up interviews used screen-recorded moments to probe participants' decision processes, acceptance/rejection of suggestions, verification practices, perceived benefits, and perceived risks. The protocol is provided in Supplementary Information 1, and de-identified response excerpts are provided in Supplementary Information 2.

Procedure and Experimental Intervention

The experiment was conducted in a controlled lab setting. Upon arrival, participants provided informed consent and completed the pre-task questionnaire. They then underwent a standardized 15-minute training and warm-up session designed to prepare them for the main task and ensure they used their assigned tools as intended.

1. **General Briefing (Both Groups):** All participants received a briefing on the research goals and instructions for the think-aloud protocol, where they were asked to verbalize their thoughts and decisions continuously. A short warm-up exercise (translating two complex sentences) was used to acclimatize them to the process.
2. **Experimental Intervention (Experimental Group):** The intervention for this group focused on priming them to use the LLM as a lexical partner, not a full translation machine. This training was crucial to isolate the effect of the "conversational lexicon." It consisted of:
    - **Framing the Task:** Participants were explicitly instructed: "For this study, please use the AI model *only* as a dynamic, conversational dictionary for exploring word choice. Do not ask it to translate full sentences or paragraphs."
    - **Demonstrating Effective Queries:** They were shown examples of effective prompts, such as, "What's a more evocative Arabic word for 'desolate' to describe a landscape?" or "What is the difference in connotation between Arabic translations for 'destin' and 'fatalité'?"
    - **Highlighting Ineffective Queries:** They were shown examples of what to avoid, primarily copy-pasting full sentences for translation.

3. **Parallel Briefing (Control Group):** To ensure fairness and control for training effects, the control group received a parallel briefing demonstrating advanced functionalities (e.g., collocations, wildcards) within their assigned tools.

Following the intervention, participants performed the 60-minute translation task. Immediately after, they completed the post-task questionnaire and participated in the semi-structured interview.

Data Analysis

Quantitative analyses were matched to the measurement structure of each outcome. Internal consistency was recalculated for the cognitive-load and creative-confidence scales. Normality and variance assumptions were checked with Shapiro–Wilk and Levene tests. Because several outcomes showed non-normality and/or unequal variances, between-group comparisons used Welch's t test with 95% confidence intervals and Hedges' g, accompanied by Mann–Whitney U sensitivity analyses.

For H1, the five pre-task self-reported GenAI-use categories were treated as repeated binary responses from the same 60 participants. Cochran's Q tested whether endorsement rates differed across functions, followed by exact McNemar comparisons with Holm adjustment. This analysis describes prior usage patterns and should not be interpreted as direct observation of the experimental interaction. For H2, final translation-quality scores were compared between the two randomized conditions using Welch's t test, with Mann–Whitney U as a robustness check. For H3, cognitive load and creative confidence were first analyzed jointly using a one-way MANOVA/Hotelling's T^2, followed by outcome-specific Welch and Mann–Whitney tests. Qualitative interview data were analyzed thematically to explain how users interacted with, verified, and sometimes rejected AI-generated lexical suggestions.

Table 1

Confirmatory and Robust Statistical Tests Used to Verify the Hypotheses

Hypothesis	Primary test	Robust / follow-up test	Key result reported in the manuscript	Decision
H1	Cochran's Q for five repeated binary pre-task GenAI-use categories	Exact McNemar comparisons with Holm adjustment; within-participant complex-vs-direct use index comparison	Q(4) = 83.90, p < .001; complex-use index > direct-use index, W = 528.00, p < .001	Supported as a pre-task usage-profile hypothesis
H2	Welch's independent-samples t-test for final product scores	Mann-Whitney U sensitivity test; Hedges' g for effect size	t(44.62) = 51.19, p < .001; U = 900.00, p < .001; Hedges' g = 13.04	Supported
H3	One-way MANOVA / Hotelling's T^2 for cognitive load and creative confidence	Follow-up Welch tests; Mann-Whitney U sensitivity tests	T^2 = 1127.92, F(2, 57) = 554.24, p < .001, Pillai's trace = .951; both follow-up outcomes p < .001	Supported

Note. *The analyses were recalculated from the current raw dataset and then written into the Results section. This table makes explicit that Cochran's Q, exact McNemar tests, Mann-Whitney U tests, and MANOVA/Hotelling's T^2 were conducted, not merely planned.*

Findings

Quantitative Data Analysis of Survey Findings

Statistical significance was set at $p < .05$. Internal consistency was high for cognitive load (Cronbach's $\alpha = .926$) and creative confidence ($\alpha = .929$). Because assumption checks indicated non-normality and unequal variances for several outcomes, Welch tests and nonparametric sensitivity analyses accompany the primary inferential results.

Hypothesis 1: Pre-Task GenAI-Use Profile

H1 concerned participants' self-reported GenAI use before the experiment, not tool behavior observed during the assigned task. Across all 60 participants, contextual explanation (n = 57, 95.0%) and synonyms or alternative phrasing (n = 56, 93.3%) were the most frequently endorsed functions, followed by single-word meaning (n = 49, 81.7%), grammar/style checking (n = 42, 70.0%), and full-sentence translation (n = 26, 43.3%).

Table 2

Frequencies and Percentages for Primary Uses of Generative AI

Query Type	Frequency (n)	Percentage (%)
Understanding the context of a difficult phrase	57	95.0
Finding synonyms or alternative phrasing	56	93.3
Looking up the meaning of a single word	49	81.7
Checking grammar and style	42	70.0
Translating entire sentences or paragraphs	26	43.3

Note. N = 60. Participants could select multiple options.

The repeated-response analysis confirmed that endorsement rates differed across the five pre-task functions, Cochran's $Q(4) = 83.90$, $p < .001$. Exact McNemar comparisons with Holm adjustment showed that contextual explanation and synonym/alternative phrasing were each endorsed more often than full-sentence translation (adjusted $p < .001$), while the two most common functions did not differ (adjusted $p = 1.000$). A within-participant comparison likewise showed a higher complex-use index (context, synonyms, grammar/style; mean proportion = .86) than direct-use index (single-word meaning and full-sentence translation; mean proportion = .63), $W = 528.00$, $p < .001$. H1 was supported as a statement about participants' prior GenAI-use profile.

Fig. 2 visualizes the pre-task usage profile. Because these items were collected before the intervention, the figure should be read as background evidence about participants' existing GenAI practices rather than as a record of experimentally observed prompting behavior.

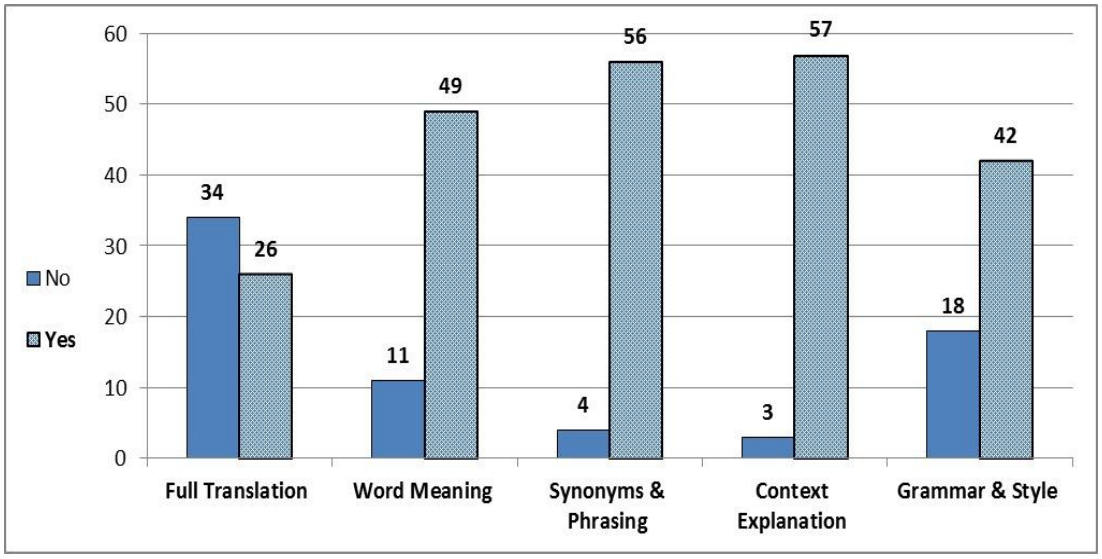


Fig. 2 Frequency of Reported Uses of Generative AI for Translation Tasks

This pattern indicates that participants entered the experiment with experience using GenAI for contextual and generative language support. It does not, by itself, establish how participants behaved during the controlled intervention; those process claims are grounded instead in the screen/audio records and interviews.

Hypothesis 2: Impact of GenAI on Final Translation Quality

H2 was supported. The Experimental Group achieved substantially higher final product scores ($M = 46.47$, $SD = 1.04$) than the Control Group ($M = 35.40$, $SD = 0.56$). Because Levene's test indicated unequal variances, Welch's test was used and confirmed a highly significant difference, $t(44.62) = 51.19$, $p < .001$. The mean difference was 11.07 points on the 50-point scale, 95% CI [10.63, 11.50], and the effect size was extremely large (Hedges' $g = 13.04$). The nonparametric Mann-Whitney U sensitivity analysis reached the same conclusion, $U = 900.00$, $p < .001$, rank-biserial $r = 1.00$.

Table 3

Advanced Comparison of Final Translation Quality by Group

Comparison	Experimental M (SD)	Control M (SD)	MD [95% CI]	Welch t(df)	p	Hedges g	Mann-Whitney U; r_rb
Final product score	46.47 (1.04)	35.40 (0.56)	11.07 [10.63, 11.50]	51.19 (44.62)	< .001	13.04	900.00; 1.00

Note. The Welch test is reported because Levene's test indicated unequal variances. CI = confidence interval; MD = mean difference; r_rb = rank-biserial correlation. The Mann-Whitney U test is included as a nonparametric sensitivity analysis.

The between-group difference in final quality scores was statistically clear in this dataset, and the rank-based sensitivity analysis yielded the same directional result. However, the effect size was extraordinarily large and the score distributions were narrow, with complete rank separation. The result should therefore be interpreted as a strong task-specific contrast between the two study conditions rather than as an estimate of the typical real-world advantage of GenAI across translation contexts.

Hypothesis 3: Effect on Cognitive Load and Creative Confidence

H3 proposed that the Experimental Group would report lower cognitive load and higher creative confidence than the Control Group. Because these two outcomes jointly describe the translator's cognitive-affective experience, they were first tested together using a one-way MANOVA/Hotelling's T^2. The multivariate group effect was significant, Hotelling's T^2 = 1127.92, $F(2, 57) = 554.24$, $p < .001$, Pillai's trace = .951, Wilks' $\lambda = .049$, indicating a very strong overall group difference in the psychological experience of the task.

Follow-up Welch tests showed that the Experimental Group reported significantly lower cognitive load (M = 7.30, SD = 0.79) than the Control Group (M = 17.03, SD = 1.47), $t(44.54) = -31.85$, $p < .001$, Hedges' g = -8.12. The Experimental Group also reported significantly higher creative confidence (M = 17.53, SD = 0.63) than the Control Group (M = 9.43, SD = 2.08), $t(34.26) = 20.42$, $p < .001$, Hedges' g = 5.21. Mann-Whitney U sensitivity tests confirmed both findings (cognitive load: U = 0.00, $p < .001$, r_rb = -1.00; creative confidence: U = 900.00, $p < .001$, r_rb = 1.00). H3 was therefore supported.

Table 4

Multivariate and Robust Tests for Cognitive Load and Creative Confidence

Outcome	Experimental M (SD)	Control M (SD)	MD [95% CI]	Welch t(df)	p	Hedges g	Mann-Whitney U; r_rb
Cognitive load	7.30 (0.79)	17.03 (1.47)	-9.73 [-10.35, -9.12]	-31.85 (44.54)	< .001	-8.12	0.00; -1.00
Creative confidence	17.53 (0.63)	9.43 (2.08)	8.10 [7.29, 8.91]	20.42 (34.26)	< .001	5.21	900.00; 1.00

Note. The one-way MANOVA/Hotelling's T^2 result was significant, $T^2 = 1127.92$, $F(2, 57) = 554.24$, $p < .001$, Pillai's trace = .951, Wilks' $\lambda = .049$. Follow-up Welch tests are reported because Levene's tests indicated unequal variances. Lower scores indicate lower cognitive load. r_rb = rank-biserial correlation.

The two self-report outcomes showed the same pronounced separation between conditions: participants in the GenAI-supported condition reported lower cognitive load and higher creative confidence. As with the product scores, the magnitude of these effects is unusually large and should be interpreted in light of the tightly controlled intervention, condition-specific training, and restricted within-group variability.

Note. Recommended confirmatory analyses include Chi-square or Fisher exact tests for each query type, Cramer's V for effect size, and a binary-logistic or GEE model to evaluate the Source Language x Query Type pattern.

Qualitative Findings: Analysis of Interview Responses

The qualitative strand drew on semi-structured interviews grounded in participants' recorded task behavior. The analysis focused on verification strategies, interaction patterns, perceived benefits, and perceived risks. The resulting themes contrast the experimental group's dialogic interaction with the control group's more static resource-switching workflow. Representative de-identified responses are available in Supplementary Information 2.

Fig. 3 is retained as a descriptive visualization of frequent terms in the interview corpus. It is not used as inferential evidence; thematic conclusions are based on coded

interview content and representative participant accounts rather than raw word frequency alone.



Fig. 3 Word Cloud of Interview Transcript Corpus

RQ1: Verification Strategies and Navigating Trust

The first qualitative research question examined how students verified AI-generated suggestions. The control group's verification process relied mainly on cross-referencing static dictionaries. By contrast, the experimental group developed a more layered process that combined AI dialogue with external validation and contextual judgment.

The first major theme was proactive verification. Participants often treated an AI suggestion as provisional and checked it against another source, real-world usage, or contextual evidence. E12, for example, described using an AI-generated phrase as a Google News search term before accepting it. A related pattern involved iterative re-prompting and self-critique: participants refined prompts, asked follow-up questions, or asked the AI to contrast candidate solutions before making a decision.

These verification practices were accompanied by critical distrust and contextual judgment. Participants recognized that grammatically fluent suggestions could still be

tonally wrong, culturally inappropriate, or insufficiently grounded in context. E4, for example, rejected a technically plausible equivalent because its connotations were too negative for the source context.

RQ2: Perceived Benefits, Risks, and the Cognitive Experience

The second qualitative research question addressed the perceived benefits and risks of the conversational lexicon. Participants' responses emphasized three connected dimensions: creative empowerment, reduced cognitive friction, and the need for professional discipline.

A prominent perceived benefit was creative and conceptual support. Experimental-group participants described the AI as a brainstorming partner that could generate alternatives, explain concepts, and help them move beyond lexical blocks. These accounts are consistent with the quantitative confidence result but do not establish an independent causal mechanism.

Participants also described reduced search fragmentation and a more continuous workflow. Several contrasted the conversational interface with moving among multiple dictionaries or websites. This qualitative pattern offers a plausible interpretation of the lower reported cognitive-load scores, while remaining a subjective account rather than a direct measure of cognitive processing.

Participants simultaneously identified risks that required discipline: dependence on fluent suggestions, over-ornamented wording, loss of stylistic voice, and premature acceptance of plausible outputs. These concerns reinforce the CLMM's regulatory layer, in which verification and final human judgment are necessary conditions for responsible use.

Table 5

Thematic Analysis of Interview Data

Theme	Codes	Example from Experimental Group (GenAI Users)	Evidence role
Creative Empowerment & Conceptual Breakthrough	*Creative partner, brainstorming, conceptual bridge, exploring options, stylistic exploration,*	"It broke through my creative wall by giving me a completely new way to think about the concept, instead	Representative qualitative theme

	overcoming blocks, idiomatic solutions	of just finding synonyms for the word 'hollow'." (E3)	
Reduced Cognitive Friction & Enhanced Flow	*Less draining, fluid process, no tab-juggling, high-level decisions, less struggle, continuous workflow, focus on writing*	"The best part... was that I never broke my 'flow state.' The task felt less like a series of hurdles and more like a continuous, fluid process of writing." (E29)	Representative qualitative theme
Proactive Verification Strategies	*Cross-referencing, external search, real-world usage, combining tools, confirmation, fact-checking*	"I used that phrase as a search term in Google News to see if Arab news outlets used it... That's my process: AI gives the creative leap, and a quick... search verifies its real-world usage." (E12)	Representative qualitative theme
Iterative Re-prompting & AI Self-Critique	*Refining prompts, follow-up questions, dialogue, clarification, asking 'why', testing the AI, comparative analysis*	"My first prompt was, 'How to describe a laugh...?' It gave me 'ضحكة صافية.' Good, but not poetic. My follow-up was, 'Now make it more literary...' It suggested 'ضحكة رنانة'." (E21)	Representative qualitative theme
Perceived Risks & the Need for Discipline	*Becoming lazy, over-reliance, loss of stylistic voice, risk of deskilling, gatekeeping, need for discipline*	"I worry about this tool making my style too... ornate... I think the real skill is learning when to use a simple word, even when the AI can give you a more complex one." (E23)	Representative qualitative theme
Critical Distrust & Contextual Judgment	*Tonally wrong, rejecting suggestions, gut feeling, plausibility check, skepticism, filtering, distrusting idioms*	"The AI's suggestion was grammatically perfect but tonally wrong for the context. It highlighted that you can't trust it blindly for nuance; you have to filter it." (E4)	Representative qualitative theme

**Weighted Percentage (WDT %) is calculated by summing the weighted percentages of the most frequent representative words for each theme as provided in the study's word frequency analysis (e.g., Creative Empowerment = creative + figurative + style + explore + concept, etc.).*

Integration of Quantitative and Qualitative Findings

Table 6 integrates the quantitative and qualitative strands. The quantitative results demonstrate measurable differences in query behavior, product quality, cognitive load, and creative confidence. The qualitative findings explain how these differences emerged: GenAI supported creative exploration, reduced process fragmentation, and encouraged new verification practices. Together, the two data strands suggest a

coherent process in which the tool's conversational affordances reshape the translator's workflow, which in turn influences both the final product and the translator's experience.

Table 7

Integration of Quantitative and Qualitative Findings

Quantitative Finding (Hypothesis)	Supporting Qualitative Theme(s)	Integrated Interpretation
H1: Pre-task GenAI-use profile: context explanation = 95.0%; synonyms/alternative phrasing = 93.3%; Cochran's Q significant.	• Creative Empowerment & Conceptual Breakthrough • Iterative Re-prompting & AI Self-Critique	The profile characterizes prior self-reported GenAI practices and should not be interpreted as observed prompting during the experiment.
H2: Significantly higher translation quality scores for the Experimental Group (M = 46.47) vs. Control (M = 35.40).	• Creative Empowerment & Conceptual Breakthrough • Critical Distrust & Contextual Judgment	The AI's ability to act as a brainstorming assistant, offering a wide range of context-aware options, empowered users to make more ambitious lexical choices. This creative exploration, guided by the translator's critical judgment, directly resulted in a final product of measurably higher quality.
H3: Significantly lower cognitive load (M = 7.30 vs. 17.03) and higher creative confidence (M = 17.53 vs. 9.43) for the Experimental Group.	• Reduced Cognitive Friction & Enhanced Flow • Proactive Verification Strategies	The conversational lexicon streamlined the research process into a single, fluid interaction. This drastic reduction in cognitive burden freed up mental resources for creative thought and verification, directly fostering the higher levels of confidence and creative agency reported by participants.

The H1 analysis describes participants' pre-task self-reported GenAI practices: contextual explanation (95.0%) and synonyms/alternative phrasing (93.3%) were more commonly endorsed than full-sentence translation. This background pattern is compatible with the conversational-lexicon concept but is analytically separate from the randomized intervention. H2 then showed a large between-condition difference in final translation quality, while the interviews illustrated how participants used AI-generated options as material for comparison, adaptation, and verification.

H3 showed lower self-reported cognitive load and higher creative confidence in the GenAI-supported condition. Interview accounts of reduced tool switching and expanded option generation provide a plausible experiential interpretation. Because the study did not experimentally isolate each interface feature, these qualitative

mechanisms should be treated as explanatory hypotheses rather than proven mediators.

Discussion

The findings show that, under a controlled task, access to a conversational LLM interface was associated with substantial differences in translation quality, perceived cognitive load, and creative confidence. The contribution is best understood as a human–AI interaction study rather than a comparison of translation engines: participants used the LLM through iterative natural-language queries, while retaining responsibility for accepting, revising, verifying, or rejecting its suggestions. This interaction pattern is consistent with the CLMM, which treats model output as one component of a user-regulated decision loop.

From Static Look-up to Dynamic Dialogue: A New Lexicographical Paradigm

The results for H1 show that students used GenAI primarily for complex lexicographical functions, especially contextual explanation and synonym/alternative phrasing, rather than for simple full-sentence translation. This finding corroborates recent work arguing that GenAI is redefining lexicographical practice and information-seeking behavior in translation [14, 27]. Traditional dictionaries and termbases support lexical retrieval, but they remain limited when students need to negotiate connotation, register, idiomaticity, and stylistic nuance. GenAI, by contrast, enables a more exploratory form of lexical inquiry in which the translator can ask follow-up questions, request alternatives, compare tones, and refine meaning through interaction.

This pattern is also consistent with studies of prompt-based translation practice, which show that LLMs are most useful when users formulate context-rich prompts rather than treating the system as a direct replacement for human translation judgment [23, 40]. In CLMM terms, the “GenAI Affordances” layer activates a new kind of “Conversational Querying and Exploration.” The student’s lexical problem is no longer resolved through a single lookup event; it becomes a sequence of inquiry, comparison, re-prompting, and selection. This explains why contextual explanation and alternative generation were dominant in the present study: these functions correspond directly to the dialogic affordances of GenAI.

The finding also aligns with broader reviews of AI-assisted translation, which suggest that AI systems are increasingly involved in translation productivity, quality support, and language-pair problem solving, while still requiring human expertise for ambiguity and contextual adaptation [30]. In Arabic-related contexts, this point is particularly important. Arabic translation often involves morphology, diglossia, culturally loaded expressions, idioms, and register-sensitive choices, all of which exceed the capacity of simple bilingual equivalence [7, 12]. The present results therefore support the view that GenAI is valuable not because it "knows" the final answer, but because it can generate a field of possible lexical options that the translator must evaluate.

CLMM and the Mediation of Lexical Decision-Making

The CLMM model helps explain the mechanism behind these findings. The model begins with an input lexical problem, such as a difficult word, phrase, idiom, register choice, or culturally marked expression. GenAI affordances then convert this problem into a mediated inquiry process: students ask for context, alternative phrasing, connotative differences, or stylistic refinement. The next layer, conversational querying, captures the empirical behavior observed in H1: students used GenAI to explore rather than simply retrieve.

The model's central claim is that the decisive stage is not AI output itself, but human-AI lexical mediation. In this stage, students generate, compare, re-prompt, refine, verify, and decide. This interpretation resonates with Vygotsky's (1978) view of tools as mediational means in cognitive development and with activity theory, which treats tools as part of a broader system of goal-directed action [18]. It also parallels distributed cognition theory, where cognition is not confined to the individual mind but distributed across people, tools, representations, and environments [24]. In the present study, lexical decision-making was distributed across the student, the AI interface, traditional resources, and the student's emerging evaluative judgment.

The CLMM also extends translation competence theory. The PACTE Group [33] conceptualizes translation competence as a multi-component ability involving bilingual, extra-linguistic, instrumental, strategic, and psycho-physiological sub-competences. The current findings suggest that AI-mediated translation requires an expanded instrumental-strategic competence: students must not only know how to use tools, but

also how to design prompts, test outputs, detect risk, and integrate AI suggestions into a coherent translation decision. This supports recent calls to treat AI literacy as a core professional competence in translation [26, 39, 45].

Creative Enhancement and Translation Quality

The results for H2 indicate that the GenAI-supported group achieved significantly higher translation-quality scores than the control group. This finding supports earlier claims that GenAI can enhance translator performance when used as an augmentative resource rather than as an uncritical replacement for human expertise [21, 25, 30]. The qualitative findings clarify why this improvement occurred: students described GenAI as a brainstorming partner that expanded their lexical repertoire, helped them overcome creative blocks, and enabled them to move beyond “safe” or literal equivalents.

This result is particularly meaningful in relation to creativity theory. Amabile’s (1996) componential model of creativity emphasizes domain knowledge, creativity-relevant processes, and motivation. In the present study, GenAI appears to have supported creativity-relevant processes by increasing the availability of alternatives and encouraging students to compare stylistic possibilities. However, the final creative act remained human: students had to decide whether the AI-generated option matched the source-text context, target-language norms, and rhetorical effect. Thus, the finding does not suggest that GenAI replaces translator creativity; rather, it provides external stimulation that may enhance creative selection when mediated by human judgment.

Arabic-English and French-Arabic translation studies further support this interpretation. Alqahtani [8] emphasizes the need for human expertise in resolving ambiguity in AI-driven Arabic translation, while Abdulmughni and Alrefaee [2] discuss the role of AI in Arabic-English literary translation in relation to linguistic accuracy and ethical implications. Similarly, Chayab and Ez-Zouaine [16] highlight the challenges of French-Arabic translation in specialized contexts, where AI assistance must be balanced with linguistic and cultural sensitivity. The present study corroborates these positions: GenAI improved translation quality not by eliminating human intervention, but by giving students more options to evaluate and refine.

Cognitive Load, Creative Confidence, and Workflow Flow

The results for H3 show that the experimental group reported lower cognitive load and higher creative confidence. This finding can be interpreted through cognitive load theory [41]. Traditional lexical searching often requires students to move between multiple dictionaries, context tools, thesauri, and search engines. This fragmented process increases extraneous cognitive load because students must manage several interfaces and compare disconnected pieces of information. GenAI reduced this fragmentation by consolidating explanation, alternative generation, and stylistic exploration into a single dialogic interface.

This does not mean that GenAI eliminates cognitive effort. Rather, it appears to redistribute cognitive effort from mechanical searching toward higher-order evaluation. Students no longer spent as much effort locating isolated equivalents; instead, they used their effort to judge appropriateness, register, idiomaticity, and stylistic fit. This interpretation aligns with Shokoohifar's (2024) focus on AI-assisted translation, cognitive load, and time, and with Aal-Hajiahmed's (2022) emphasis on the cognitive processes and strategies involved in English-Arabic and Arabic-English interpreting. The present study extends this line of work by showing that AI-mediated lexical support may reduce process friction while preserving, or even intensifying, the need for strategic judgment.

The increase in creative confidence also supports the CLMM model. In the model, positive outcomes emerge when GenAI affordances are filtered through the critical regulatory layer. Students felt more confident not simply because AI gave them answers, but because it helped them generate options, test alternatives, and justify decisions. This explains why lower cognitive load and higher creative confidence appeared together: when lexical exploration became more fluid, students were able to devote more attention to creative and evaluative decisions.

Verification, Risk Filtering, and AI Literacy

A central contribution of this study is its demonstration that the benefits of GenAI depend on verification. Participants did not simply accept AI-generated suggestions. They cross-checked

outputs, used AI suggestions as search terms, asked follow-up questions, requested self-critique, and relied on contextual judgment. This finding strongly supports the CLMM's "Critical Regulatory Layer," which positions AI literacy, contextual judgment, cross-checking, and final decision-making as necessary mediators between AI affordances and translation outcomes.

This finding is consistent with Krüger's (2023) argument that AI literacy is essential for the language industry and with Martín's (2024) discussion of the risks of AI-generated terminological definitions. It is also reinforced by Al-Jarf's (2025) warning that AI may fail to answer specific linguistic questions accurately. The present study confirms this concern at the level of student practice: participants recognized that AI-generated expressions could be grammatically fluent but tonally wrong, culturally inappropriate, or stylistically excessive. Their responses therefore support the view that AI-generated lexical suggestions should be treated as hypotheses to be tested, not solutions to be accepted automatically.

The risk filter in CLMM captures this tension. GenAI may produce hallucinations, lexical smoothing, over-generic phrasing, tonal mismatch, or culturally unsuitable expressions. Previous studies have identified similar limitations in Arabic-English AI translation, including ambiguity, code-switching, legal terminology, and cultural nuance [8, 9, 10]. The present findings corroborate this literature by showing that students were aware of these risks and developed practical strategies to manage them. Their behavior supports a co-regulatory view of AI use: the machine generates possibilities, but the translator regulates, filters, and authorizes the final lexical decision.

Source Language, Cross-Linguistic Moderation, and

The revised The available data did not include a directly coded categorical source-language variable, so the study could not validly test whether English-Arabic and French-Arabic translation produced different query profiles. The exploratory proficiency analysis did not show significant associations with query types, but proficiency indicators are not equivalent to source-language/department membership. Therefore, the study should avoid claiming that

Nevertheless, the CLMM model provides a useful theoretical basis for future testing of The model treats source language, linguistic distance, idiom, register, genre, and cultural reference as moderating conditions. Prior work suggests that these factors

are likely to shape AI-mediated translation behavior. English-Arabic translation may intensify problems of idiom, cultural meaning, morphology, and diglossia [7, 12], while French-Arabic translation may involve distinctive challenges of register, formal nuance, and stylistic precision [16]. In specialized domains such as legal or medical translation, these effects may become even more pronounced because terminological accuracy and ethical responsibility are central [10, 19].

Thus, although Future analysis should code source language directly and examine whether English-Arabic and French-Arabic students differ in their use of contextual explanation, synonym generation, grammar/style checking, idiom clarification, and full-sentence translation. Such analysis would allow the CLMM moderation layer to be tested more directly.

Implications of the CLMM for Translation Pedagogy

The findings suggest that translator training should move beyond teaching students whether AI is useful or risky. The more important pedagogical question is how students can regulate AI use. The CLMM model offers a pedagogical sequence for this purpose: identify the lexical problem, formulate a context-rich prompt, compare generated alternatives, cross-check outputs, request clarification or self-critique, evaluate risk, and make a final human decision. This sequence transforms AI use from a shortcut into a structured learning activity.

This implication is consistent with Vygotsky's (1978) emphasis on mediated learning and with activity-theoretical approaches that view tools as embedded in social and educational practices [18]. In translation classrooms, GenAI can function as a scaffold for lexical exploration, but only if students are trained to question its outputs. The goal should not be blind fluency, but reflective lexical agency. This is especially important in Arabic translation contexts, where plausible AI outputs may still fail at the level of idiomaticity, cultural resonance, religious reference, or register.

Summary

Overall, H1–H3 support a bounded version of the study's central argument. Participants reported prior use of GenAI for contextual and generative lexical functions, and the randomized tool condition produced strong task-specific contrasts in product

and self-report outcomes. The qualitative strand suggests that these contrasts co-occurred with iterative prompting, option comparison, external verification, and contextual filtering. The findings therefore support studying conversational LLMs as interactive cognitive tools whose value depends on user regulation rather than treating them as autonomous substitutes for human judgment.

Limitations

Several limitations constrain interpretation. First, the sample comprised senior undergraduate translators, so the findings may not generalize to professional users. Second, the tasks were short literary excerpts translated into Arabic; technical, legal, or high-stakes domains may produce different interaction patterns. Third, the intervention bundled access to an LLM with condition-specific training and prompting instructions, so the experiment estimates the effect of the configured workflow rather than the isolated effect of a model or interface feature. Fourth, the outcome distributions were unusually compressed and showed near-complete or complete group separation, producing very large standardized effects; replication with broader tasks, blinded scoring archives, and more heterogeneous samples is essential. Fifth, the available analysis archive lacks row-level source-language coding and rater-level scores, preventing direct analysis of language-pair moderation and independent recalculation of inter-rater reliability. Finally, rapid model change limits direct generalization to future LLM versions.

Conclusion

This study examined generative AI as a conversational interface for human lexical decision-making. In a randomized comparison, the GenAI-supported group achieved higher evaluated translation scores and reported lower cognitive load and higher creative confidence than the traditional-resource group. Before the task, participants also reported using GenAI mainly for context explanation and alternative phrasing rather than full-sentence translation. Interviews showed that productive interaction depended on iterative prompting, comparison of alternatives, cross-checking, and contextual judgment.

The quantitative results support three specific conclusions within the study design: pre-task GenAI use was oriented toward contextual and generative functions; the GenAI-supported condition outperformed the control condition on the 50-point product score; and the GenAI-supported condition reported a more favorable cognitive-affective profile. The very large effect sizes should not be generalized as typical population effects because the study used a tightly controlled intervention and observed restricted within-group dispersion.

The CLMM provides a useful interaction-level interpretation of these findings. A lexical problem initiates a natural-language query; the model expands the option space; the user compares and tests alternatives; and verification and contextual judgment regulate the final decision. This sequence keeps agency with the user and locates the computational contribution in interactive generation and explanation rather than autonomous task replacement.

The interviews show that participants did not uniformly accept fluent outputs. They cross-checked suggestions, refined prompts, requested explanations, and rejected options that were culturally or tonally inappropriate. These behaviors identify concrete design and training requirements for human-centered GenAI systems: interfaces should support comparison, provenance-aware verification, uncertainty-sensitive prompting, and easy movement between generation and external evidence.

The study contributes to research on human–AI interaction, natural-language interfaces, and AI-supported knowledge work by showing how a conversational model can function as an augmentative lexical decision aid. Its value in this setting derived from interactive exploration rather than automatic translation. The results also underscore that fluent generation is not sufficient: human verification, trust calibration, and final decision authority remain central to safe and effective use.

Recommendations

These findings suggest practical implications for future research, translator training, and the translation industry.

For Future Research

Future work should replicate the design with professional translators, larger and more heterogeneous samples, multiple LLMs, and tasks from domains with different risk profiles. Studies should preserve row-level source-language and rater-level data, preregister primary outcomes, and experimentally separate interface effects from training effects. Interaction logs could also be modeled at the prompt–response level to test whether verification, prompt refinement, latency, or option diversity mediate performance and cognitive-load outcomes.

For Translator Training

Training should extend beyond tool access to interaction literacy: formulating context-rich prompts, comparing alternatives, checking real-world usage, recognizing fluent but unsuitable outputs, and maintaining responsibility for final decisions.

For the Translation Industry

For applied language-technology workflows, the findings support human-in-the-loop deployment rather than unmonitored substitution. Organizations should pair GenAI access with verification protocols, domain-specific guidance, and audit practices appropriate to the consequences of error.

Ethics approval and consent to participate

The study involved human participants and was conducted in accordance with the Declaration of Helsinki. Ethical approval was granted by the Faculty of Education Ethics Committee, Beni-Suef University, Egypt (Approval No. FoE-002-02-25). Written informed consent was obtained from all participants before data collection.

Consent for publication

The manuscript contains only de-identified participant quotations. Participants provided written consent for the research use of their questionnaire responses, recordings, and interview data. No directly identifying personal information is reported.

Competing interests

The authors declare that they have no competing financial or non-financial interests relevant to this work.

Funding

This research received no external funding.

Data availability

De-identified quantitative data, statistical-output files, the validated instruments, and de-identified interview-response materials are supplied as Supplementary Information with this submission. Raw screen/audio recordings are not publicly shared because they may contain identifiable participant information. The supplied quantitative archive supports the analyses reported here; however, it does not contain row-level source-language coding or rater-level scores.

Author contributions

M.M.: conceptualization, formal analysis, methodology, writing—original draft, and writing—review and editing. F.A.: methodology, qualitative analysis, investigation, and writing—review and editing. Both authors contributed to data collection, reviewed the final manuscript, and approved the submitted version.

Supplementary Information

Supplementary Information 1: validated pre-task/post-task questionnaires and semi-structured interview protocol. Supplementary Information 2: de-identified interview responses. Supplementary Data 1: de-identified raw quantitative dataset. Supplementary Data 2–5: legacy hypothesis-specific SPSS output files supplied for transparency. The manuscript uses the recalculated analyses reported in the Results section as the primary inferential record.

References

[1] Aal-Hajiahmed, M. J. (2022). Cognitive processes in simultaneous interpreting from English into Arabic and from Arabic into English: A study of problems and interpreter strategies [Doctoral dissertation, Universitat Autònoma de Barcelona]. TDX.

[2] Abdulmughni, H. M. A., & Alrefaee, S. M. A. S. (2025). Artificial intelligence in Arabic-English literary translation: Technological innovations, linguistic accuracy, and ethical implications. International Journal of Language and Literary Studies, 7(3), 107–129. https://doi.org/10.36892/ijlls.v7i3.2120

[3] Ait Elfqih, K. (2025). Addressing machine translation of legal terminology in Arabic legislation: From error evaluation to systems optimisation [Doctoral dissertation, L'Orientale University of Naples]. Unior Open Archive. https://www.unora.unior.it/handle/123456789/2237

[4] Aladrović, J. (2024). Comparative analysis of MemoQ Adaptive Generative Translation (AGT) with conventional machine translation (MT) engines [Master's thesis, Josip Juraj Strossmayer University of Osijek]. FFOS-repozitorij. http://repozitorij.ffos.hr/id/eprint/6122

[5] AlAfnan, M. A. (2025). Artificial intelligence and language: Bridging Arabic and English with technology. Journal of Ecohumanism, 4(1), 240–256. https://doi.org/10.62754/joe.v4i1.4961

[6] Al-Jarf, R. (2025). 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

[7] Almanea, M. (2024). Deep learning in written Arabic linguistic studies: A comprehensive survey. IEEE Access, 12, 25176–25203. https://doi.org/10.1109/ACCESS.2024.3364239

[8] Alqahtani, D. A. (2024). Transcending ambiguities: Enhancing AI-driven Arabic to English translations with human expertise. Journal of Languages and Translation, 11(3), 59–81. https://doi.org/10.21608/jltmin.2024.373620

[9] Alqarni, M. D. (2025). Investigating the translation dynamics of Arabic-English code-switching by AI and EFL university students in Saudi Arabia. International Journal of Linguistics, Literature and Translation, 8(6), 72–84. https://doi.org/10.32996/ijllt.2025.8.6.7

[10] Alwazna, R. Y., Hurley, Z., & Elyas, T. (2025). Integrating technology in translating laws from and into Arabic: Benefits, implications and limitations. International Journal for the Semiotics of Law/Revue internationale de Sémiotique juridique, 38(6), 1851–1862. https://doi.org/10.1007/s11196-025-10295-8

[11] Amabile, T. M. (1996). Creativity in context: Update to the social psychology of creativity. Westview Press.

[12] Azizov, D. (2024). From idioms to algorithms: Translating culture-specific expressions in AI systems. Iconic Research and Engineering Journals, 7(9), 41–48.

[13] Baya, H., & Becheri, R. (2024). Investigating the latest technologies in the translation industry and their impact on EFL learning [Master's thesis, Ibn Khaldoun University of Tiaret]. DSpace. http://dspace.univ-tiaret.dz/handle/123456789/13769

[14] Cai, Y., & Tian, S. (2025). Student translators' web-based vs. GenAI-based information-seeking behavior in translation process: A comparative study. Education and Information Technologies. https://doi.org/10.1007/s10639-024-12791-8

[15] Chang, C. C., Li, C. F., Lee, C. H., & Lee, H. S. (2025). Enhancing low-resource minority language translation with LLMs and retrieval-augmented generation for cultural nuances. arXiv. https://doi.org/10.48550/arXiv.2405.10829

[16] Chayab, F., & Ez-Zouaine, Y. (2026). Integrating social AI into French–Arabic environmental translation: Common challenges and solutions. In Transforming humanities and social sciences through AI (pp. 367–390). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-3436-3.ch013

[17] Cruz, C. J. X. (2024). Transforming competition into collaboration: The revolutionary role of multi-agent systems and language models in modern organizations. arXiv. https://doi.org/10.48550/arXiv.2403.07769

[18] Engeström, Y. (1987). Learning by expanding: An activity-theoretical approach to developmental research. Orienta-Konsultit.

[19] Gatsiou, C. (2024). Artificial intelligence (AI) in translation and interpreting: Exploring AI's potential in translation and interpreting profession and pedagogy in

the medical context and its ethical implications [Master's thesis, National and Kapodistrian University of Athens]. Pergamos Digital Library.

[20] Ghali, M. K. (2024). Leveraging generative AI and in-context learning to reshape human-text interaction: A novel paradigm for information retrieval, named entities extraction, and question answering [Doctoral dissertation, École de technologie supérieure]. ProQuest Dissertations & Theses Global.

[21] Guerberof, A., & Asimakoulas, D. (2024). Creative skills development: Training translators to write in the era of AI. Hermes: Journal of Language and Communication in Business, 65, 125–144. https://doi.org/10.7146/hjlcb.v65i.143249

[22] Harvey, W. (2024). How artificial intelligence is improving human communication with the processing of natural language. EPH: International Journal of Science and Engineering, 9(3), 1–9. https://doi.org/10.53555/ephijse.v9i3.2757

[23] He, S. (2024). Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. arXiv. https://doi.org/10.48550/arXiv.2403.00127

[24] Hutchins, E. (1995). Cognition in the wild. MIT Press. https://doi.org/10.7551/mitpress/1881.001.0001

[25] Kornacki, M., & Pietrzak, P. (2024). Hybrid workflows in translation: Integrating GenAI into translator training. In A. Pym & Y. Hao (Eds.), How to augment language skills: Generative AI and machine translation in language learning and translator training (pp. 95–109). Routledge.

[26] Krüger, R. (2023). Artificial intelligence literacy for the language industry: With particular emphasis on recent large language models such as GPT-4. Lebende Sprachen, 68(3), 365–393. https://doi.org/10.1515/les-2023-0020

[27] Lew, R. (2024). Dictionaries and lexicography in the AI era. Humanities and Social Sciences Communications, 11(1), 1–8. https://doi.org/10.1057/s41599-023-02539-x

[28] Martikainen, M. (2024). Estimating generative AI impacts in public social and health care language translation services [Master's thesis, LUT University]. LUTPub. https://lutpub.lut.fi/handle/10024/167733

[29] Martín, A. S. (2024). What generative artificial intelligence means for terminological definitions. arXiv. https://doi.org/10.48550/arXiv.2402.16139

[30] Mohamed, Y. A., Khanan, A., Bashir, M., Mohamed, A. H. H., Adiel, M. A., & Elsadig, M. A. (2024). The impact of artificial intelligence on language translation: A review. IEEE Access, 12, 25553–25579. https://doi.org/10.1109/ACCESS.2024.3366802

[31] Moorkens, J., Way, A., & Lankford, S. (Eds.). (2024). Automating translation. Routledge.

[32] Naveen, P., & Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. iScience, 27(3), Article 109156. https://doi.org/10.1016/j.isci.2024.109156

PACTE Group. (2003). Building a translation competence model. In F. Alves (Ed.), *Triangulating translation: Perspectives in process-oriented research* (pp. 43–66). John Benjamins. https://doi.org/10.1075/btl.45.06pac

[33] Puppel, M. (2024). A comparative study of technical and creative text translation: Evaluating the performance of ChatGPT [Master's thesis, University of Malta]. OAR@UM. https://www.um.edu.mt/library/oar/handle/123456789/122718

[34] Pym, A., & Hao, Y. (Eds.). (2024). How to augment language skills: Generative AI and machine translation in language learning and translator training. Routledge.

[35] Román Martínez, J., Triana Robles, D., & Durán Muñoz, I. (2025). Generative artificial intelligence and machine translators in Spanish translation of early vulnerability cybersecurity alerts. Applied Sciences, 15.

[36] Saoudi, Y., & Gammoudi, M. M. (2023). Trends and challenges of Arabic chatbots: Literature review. Jordanian Journal of Computers and Information Technology, 9(3), 253–274. https://doi.org/10.5455/jjcit.71-1672322440

[37] Shi, J., Jain, R., Doh, H., Suzuki, R., & Ramani, K. (2023). An HCI-centric survey and taxonomy of human-generative-AI interactions. arXiv. https://doi.org/10.48550/arXiv.2311.07223

[38] Shokoohifar, M. (2024). Artificial intelligence-assisted translation: A study of cognitive load and time through EEG [Doctoral dissertation, Allameh Tabataba'i University].

[39] Shreve, G. M. (2024). Competences for the human-machine era in translation. In A. Pym & Y. Hao (Eds.), How to augment language skills: Generative AI and machine translation in language learning and translator training (pp. 110–123). Routledge.

[40] Siu, S. C. (2023). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. SSRN. https://doi.org/10.2139/ssrn.4448091

[41] Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

[42] Tekwa, K. (2024). Artificial intelligence, corpora, and translation studies. In M. Ji & S. Oakes (Eds.), The Routledge handbook of corpus translation studies. Routledge.

[43] Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

[44] Weigang, L., & Brom, P. C. (2025). LLM-BT-TERMS: Back-translation as a framework for terminology standardization and dynamic semantic embedding [Unpublished manuscript]. ResearchGate. https://www.researchgate.net/publication/380720610

[45] Zimina-Poirot, M., & Gledhill, C. (2025). Towards clear, simple, and inclusive legal language: Exploring generative artificial intelligence (GAI) models as content rephrasing assistants in multilingual legal drafting. International Journal of Language & Law, 14.

[46] Zulianti, I. (2024). Navigating the evolving lexicology landscape: A systematic literature review. ELite Journal: International Journal of Education, Language, and Literature, 4(1), 162–176. https://doi.org/10.33758/elite.v4i1.1398

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FinalInstrumentsValidated.docx
- Hypotehsis2dataset2.xlsx
- Hypothesis3dataset3.xlsx
- Hypothesis4dataset4.xlsx
- Hypothesis1dataset1.xlsx
- RAWDATA.xlsx
- SPSSOutput.sav
- Dataset2.docx
- WordFrequencyQuery.xlsx