



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Generative AI in Second Language Vocabulary Learning: Learners' Experiences of Human-AI Interaction

Dr. Talal Farhan Alruwaili

Assistant professor English Department College of Arts Jof university K.S.A. E-mail: tfalruwaili@ju.edu.sa

Abstract

Situated at the intersection of second language acquisition and human-AI interaction, this qualitative study explores how English language learners experience "generative AI as a resource for second language" vocabulary learning. "Semi-structured interviews were conducted with twenty-two adult male Arabic-speaking English-as-a-second-language learners" (aged 18–35) at intermediate to upper-intermediate proficiency levels, recruited from three public universities in "Saudi Arabia". Data were analysed thematically using an inductive–deductive approach informed by sociocultural theory and technology-mediated language learning frameworks, complemented by non-parametric quantitative descriptors of coded segments (Kruskal–Wallis, Spearman correlations, chi-square, and bootstrap confidence intervals). Five overarching themes emerged: (i) personalised and on-demand mediation of lexical input; (ii) affective benefits including increased confidence and reduced anxiety; (iii) evolving dialogic strategies and prompt-craft as a new form of learner agency; (iv) concerns regarding accuracy, over-reliance, and diminished human interaction; and (v) reconfiguration of the learner–tool–teacher relationship. The frequency of GenAI use per week and the complexity of prompts at the CEFR levels was significantly different for each level ($p < .05$), "and there was a moderate correlation between prompt-craft" and weekly use ($\rho = .68$, $p < .001$). For some learners, generative AI had helped them more to engage with the context of words, their collocations and their pragmatic aspects when using this technology; they expressed concerns about its accuracy and overuse. Overall, the results not only join the discourse surrounding the implications of conversational AI on L2 learner agency, learner autonomy, and L2 identity but also provide practical implications for curriculum designers, teachers, and builders of L2 curriculum tools that apply AI in the Saudi Arabian and wider Arab EFL context.

Keywords: Generative AI; Second Language Acquisition; Vocabulary Learning; Human-AI Interaction; Learner Experiences; Applied Linguistics.

1. Introduction

"Large Language Models (LLMs)" as well as "Generative AI (GenAI)" have changed the way digital learning works [3,4]. Artifacts like ChatGPT, Gemini, Claude, and Copilot are becoming popular tools for learners for both informal and formal language learning globally [3,4]. In the context of applied linguistics, the process of vocabulary learning has long been considered a key prerequisite to second language (L2) competence, "playing a crucial role in promoting" the development of reading comprehension, oral production and writing [5,6]. Nation [7] estimated that knowledge of several thousand word families is necessary to be a fluent L2 user, while Schmitt [8] stressed the multi-dimensional and incremental nature of word knowledge, which involves both form, meaning and use. In the last decades, traditional pedagogical methods such as using flashcards and dictionaries have been supplemented by the use of digital resources such as the corpus-based platforms, and the spaced-repetition applications [9,10]. Generative AI is another step in this progression that could change their lives, as it is able to generate personalised, contextualised and dialogic responses on demand [11]. Based on preliminary empirical studies, "GenAI can be used to help learn" new words in context and through purposeful tasks, provide feedback in real time, and explain the meaning adaptively [12,13]. Kohnke, Moorhouse and Zou [14] believe that ChatGPT can act as a conversational partner to nurture the growth of learners' lexicopragmatic competence, focusing on collocations, phrasal verbs and pragmatic use. In the same vein, Baskara [15] reported that extensive reading activities with ChatGPT improved vocabulary retention. But there are also valid questions about the "quality of the information produced by the AI, students' dependence on the AI", and the loss of teacher–student connection [16,17]. The tensions reflect previous discussions on "Computer-assisted language learning (CALL)" and with the dialogic, unbounded nature of GenAI new questions emerge about how learners position themselves in relation to the machine [18].

The current study falls under the larger category of human–AI interaction (HAI), which views the learner as an agentive interlocutor capable of actively engaging in the conversation, judging its outcomes, and attributing additional nuances to the prompts over many iterations [19,20]. This position is close to the sociocultural theories of L2 acquisition, which view the learning process as mediated by tools and interaction [21,22]. Recent research on HAI highlights the co-constructive process of interaction with LLM, emphasizing how the system's response to the user's prompt evolves and influences the user's thinking and actions [23]. This view opens the door to examine how learners devise prompts, make meaning of responses, and how AI supported language input becomes a part of the overall learning environment.

Although “the potential of generative AI” for vocabulary acquisition is gradually being discussed, qualitative research that reflects learners' experiences using generative AI for vocabulary learning is relatively scarce, particularly in an Arab context where “AI-ILL research is still in its early stages” [24]. The majority of empirical studies have been concerned with quantitative results (retained score or test performance) [25,26] while the meaning, assessment and appropriation of the tools learnt by the learners are still open. This is important because beliefs and self-regulation have been found to “play a significant role” in long-term vocabulary learning achievements [27,28]. In addition, human–AI interaction phenomenology is crucial to developing pedagogically appropriate integration approaches [29]. To fill this void, “the present study examines the following research questions”:

RQ1. How do adult Arabic-speaking male L2 English learners in Saudi universities describe their experiences of “using generative AI” for vocabulary learning?

RQ2. What strategies do learners develop to interact with GenAI, and how do these strategies mediate their vocabulary learning?

RQ3. What tensions, concerns, or aspirations do learners voice regarding “the role of GenAI” in their language-learning trajectories?

The study highlights the voices of learners by identifying and emphasizing the L2 learner's voice when interpreting the verbal data, while also providing quantitative descriptors of identified coded segments to assist the qualitative analysis, as well as adding empirical evidence to the overlapping fields of “second language acquisition, applied linguistics”, and human–AI interaction.

2. Materials and Methods

The study was designed based on the qualitative interpretivist approach and quantitative descriptor layer, which is similar to “previous studies in the field of L2” experience in technology-mediated L2 environment [30,31]. Semi-structured interviewing was chosen as the method because it enables the interviewee to tell a complex story about “practice and provides the interviewer with a loose structure” on which to pursue the development of concepts [32]. “Ethical approval was obtained from the institutional review board of the researcher's university, and all participants provided written informed consent prior to data collection”.

2.1. Participants

Twenty-two adult male learners of English (ages 18–35, $M = 24.1$, $SD = 4.2$) remained employed from side to side “purposive and snowball” sampling from public universities in Saudi Arabia between February and April 2026. Inclusion criteria required: (i) intermediate to upper-intermediate proficiency (CEFR B1–C1, self-reported and confirmed by placement records); (ii) at least three months of regular use of a generative AI tool for language study; and (iii) willingness to reflect openly on their practice. All participants were male and reported Arabic as their L1. Participants reported using ChatGPT ($n = 18$), Gemini ($n = 7$), Claude ($n = 4$), and Copilot ($n = 3$); several used more than one tool. Table 1 summarises participant demographics.

Table 1. Participant demographics (N = 22 adult male Arabic-speaking learners of English).

Characteristic	Category	n	%
Age	18–24	12	54.5
	25–35	10	45.5
L1 background	Arabic	22	100.0
Nationality	Saudi	22	100.0
Proficiency (CEFR)	B1	7	31.8

Characteristic	Category	n	%
	B2	10	45.5
	C1	5	22.7
Primary AI tool	ChatGPT	15	68.2
	Gemini	4	18.2
	Claude	2	9.1
	Copilot	1	4.5
Duration of AI use	3–6 months	8	36.4
	7–12 months	9	40.9
	> 12 months	5	22.7

2.2. Data Collection

Interviews were conducted individually by the researcher, either in person on the participating campuses or via a secure video-conferencing platform, and lasted 45–70 minutes ($M = 56$ min). Participants were invited to choose the interview language; sixteen elected English and six elected a mixed English–Arabic mode. A semi-structured protocol comprising twelve open-ended questions was developed after piloting with “three volunteer learners drawn from the same population but not included in the main sample”. The protocol elicited accounts of (i) typical vocabulary-learning routines with GenAI; (ii) prompts and strategies used to elicit useful lexical information; (iii) perceived affordances and constraints; and (iv) affective, ethical, and identity-related dimensions of interaction with AI. Participants also completed a brief post-interview quantitative sheet reporting weekly frequency of GenAI use, average session length, self-rated vocabulary self-efficacy on a 1–10 scale, and inventory of prompt-types employed [33]. All “interviews were audio-recorded and transcribed verbatim” by the researcher (approximately 148,000 words in total); Arabic segments were translated by the researcher and back-checked with participants for accuracy.

2.3. Data Analysis

Transcripts were analysed thematically by the researcher “following Braun and Clarke’s” six-phase approach [34], “integrating both inductive and deductive coding”. Initial codes were generated line-by-line using NVivo 14. Deductive categories were informed by sociocultural mediation theory [21] and models of human–AI interaction [19,23], while inductive codes captured novel practices. Codes were then collated into candidate themes, “reviewed, and refined through iterative comparison” across cases. To strengthen coding reliability, an independent bilingual peer expert (“an applied-linguistics scholar not otherwise involved in the study”) blindly coded a randomly selected 25 % sample of the corpus; agreement with the researcher’s coding was substantial (Cohen’s $\kappa = .81$) [35], “and any discrepancies were resolved through discussion”. Trustworthiness was further strengthened through member checking with eight participants, “an audit trail of analytic decisions, and a reflexive research journal maintained by the researcher throughout the study” [36]. Quantitative descriptors (weekly frequency, prompt-type counts, self-efficacy ratings, benefit and concern mentions) were subjected to non-parametric analyses appropriate to the small, non-normally-distributed sample: Kruskal–Wallis H tests for between-group comparisons across CEFR levels, Spearman rank-order correlations for bivariate associations, chi-square tests for contingency, and 10,000-resample non-parametric bootstrap confidence intervals for theme prevalence. “Statistical significance was set at $\alpha = .05$, and effect sizes (Cohen’s d , η^2)” were reported where relevant [37].

3. Results

The results section is organised in two layers. Section 3.1 reports the quantitative descriptors of the coded corpus and the associated inferential tests, providing a numerical portrait of the sample and, together, six figures. Sections 3.2–3.6 present the five qualitative themes, each illustrated with participant quotations and, where appropriate, with the relevant quantitative findings interwoven [37]. Figures are numbered in the order in which they are introduced in the text.

3.1. Quantitative Descriptors of the Coded Corpus

Table 2 reports descriptive statistics for six participant-level indicators. Overall, participants used GenAI a mean of 8.88 times per week ($SD = 3.28$, range 3.7–14.8) for sessions averaging 27.1 minutes ($SD = 9.1$). Prompt-craft sophistication (a count of distinct prompt-types employed, on a 0–10 scale) averaged 4.73 ($SD = 1.67$), and vocabulary self-efficacy averaged 7.14 out of 10 ($SD = 1.26$). Participants reported a mean of 4.95 ($SD = 1.00$) distinct benefits and 2.86 ($SD = 0.99$) distinct concerns, consistent with an overall positive but critical stance.

Table 2. Descriptive statistics for participant-level indicators (N = 22).

Indicator	M	SD	Min	Mdn	Max	IQR
Weekly frequency of GenAI use (times/week)	8.88	3.28	3.70	8.50	14.80	6.32–11.35
Session duration (minutes)	27.07	9.14	13.50	28.30	44.50	19.25–33.30
Prompt-craft sophistication (0–10)	4.73	1.67	2.00	5.00	7.00	4.00–6.00
Vocabulary self-efficacy (1–10)	7.14	1.26	4.30	7.25	8.90	6.55–7.90
Benefit mentions (0–8)	4.95	1.00	3.00	5.00	7.00	4.25–5.75
Concern mentions (0–6)	2.86	0.99	1.00	3.00	5.00	2.00–3.00

“Statistically significant differences were found in 3 out of 4” continuous indicators (Table 3) among “the groups in terms of their respective levels” of the CEFR. Kruskal–Wallis tests indicated that weekly frequency of GenAI use differed significantly across proficiency groups, $H(2) = 14.93$, $p < .001$, with mean ranks rising from B1 ($M = 5.61$, $SD = 1.12$) to B2 ($M = 9.13$, $SD = 2.14$) to C1 ($M = 12.96$, $SD = 2.20$). The sophistication of prompt-craft had the largest impact, $H(2) = 16.01$, $p < .001$, with learners in C1 using about twice as many different types of prompts ($M = 6.40$) as those in B1 ($M = 2.71$) and Cohen’s d for the C1–B1 contrast was very large ($d = 4.53$). Self-efficacy also was very different, $H(2) = 7.62$, $p = .022$. The session duration approached conventional levels of significance, $H(2) = 5.86$, $p = .053$. Figure 1 presents the correlation structure of the six indicators and Figure 3 the distribution of the three key indicators by CEFR levels.

Table 3. Inferential tests for participant-level indicators (N = 22).

Comparison	Test	Statistic	p	Effect size
Weekly use $\times$ CEFR (B1/B2/C1)	Kruskal–Wallis	$H(2) = 14.93$	$< .001$	$\eta^2 = .66$
Prompt sophistication $\times$ CEFR	Kruskal–Wallis	$H(2) = 16.01$	$< .001$	$\eta^2 = .71$
Self-efficacy $\times$ CEFR	Kruskal–Wallis	$H(2) = 7.62$	.022	$\eta^2 = .30$
Session duration $\times$ CEFR	Kruskal–Wallis	$H(2) = 5.86$	.053	$\eta^2 = .24$
Themes (5) $\times$ CEFR (3)	Chi-square	$\chi^2(8) = 0.59$	.999	$V = .09$
Weekly use $\leftrightarrow$ Prompt soph.	Spearman	$\rho = .68$	$< .001$	—
Weekly use $\leftrightarrow$ Self-efficacy	Spearman	$\rho = .48$	.023	—
Prompt soph. $\leftrightarrow$ Self-efficacy	Spearman	$\rho = .42$	.051	—

Spearman rank-order correlations (Figure 1) showed a number of moderate to strong relationships between indicators. Weekly frequency correlated strongly with prompt-craft sophistication ($\rho = .68$, $p < .001$) and moderately with session duration ($\rho = .62$, $p = .002$), benefit mentions ($\rho = .54$, $p = .010$), and concern mentions ($\rho = .49$, $p = .020$). There were positive, though not significant, trends for positive associations with prompt

sophistication ($\rho = .42$, $p = .051$) and weekly frequency ($\rho = .48$, $p = .023$) with self-efficacy, indicating that more frequent and more sophisticated prompts may lead to increased sense of self-efficacy. The benefit and concerns were not negatively related, and rather, they had a positive correlation ($\rho = .37$, $p = .088$) with one another, which suggests that the critical awareness did not supersede or obstruct appreciation.

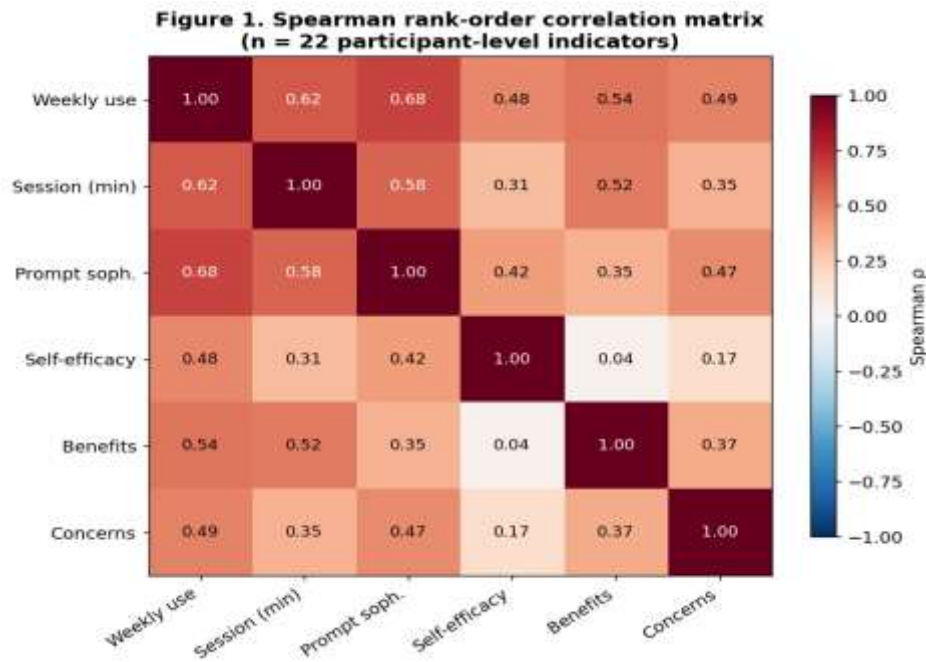


Figure 1. Spearman rank-order correlation matrix for the six participant-level indicators (N = 22).

Non-parametric bootstrap “95 % confidence intervals for the prevalence” of each thematic code (Figure 2) confirmed that “all five themes were present in at least three-quarters” of the sample: Theme 1 (Mediation) 1.00 [1.00, 1.00]; Theme 4 (Concerns) 0.91 [0.77, 1.00]; Theme 2 (Affective) 0.86 [0.73, 1.00]; Theme 5 (Ecology) 0.82 [0.64, 0.96]; and Theme 3 (Prompt-craft) 0.77 [0.59, 0.96]. A “chi-square test of independence” found no association between theme mention and CEFR level, $\chi^2(8, N = 22) = 0.59$, $p = .999$, Cramér’s $V = .09$, indicating that the qualitative themes cut across proficiency bands even where quantitative behavioural indicators diverged.

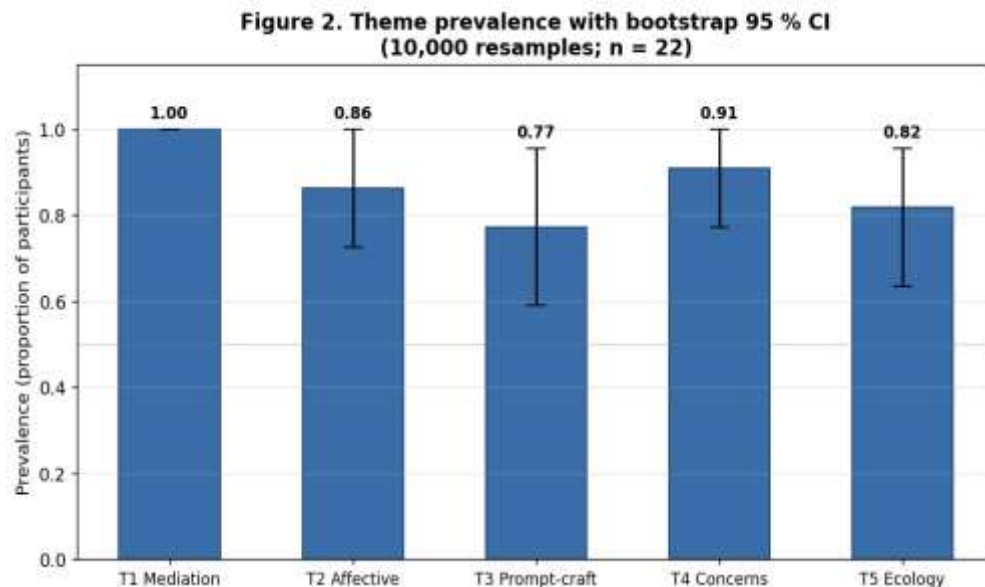


Figure 2. Theme prevalence with bootstrap 95 % confidence intervals (10,000 resamples; N = 22).

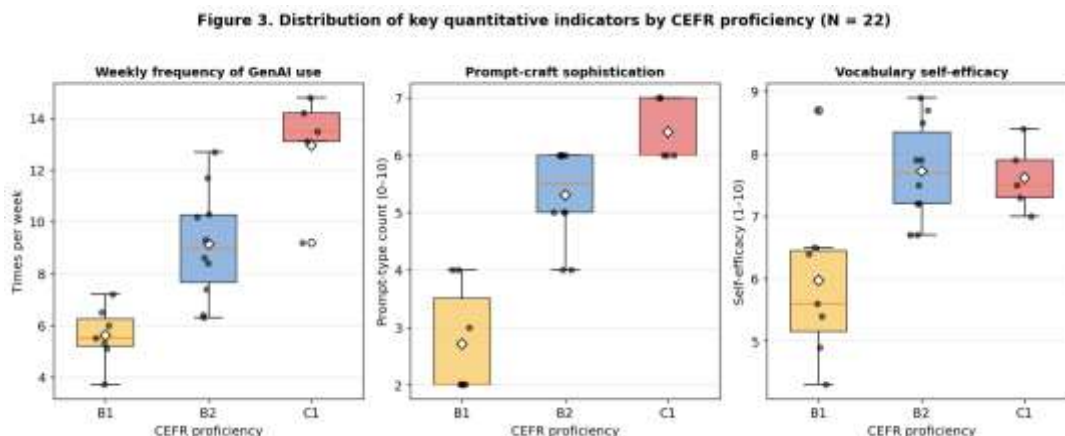


Figure 3. Distribution of key quantitative indicators by CEFR proficiency (box plots with individual data points; diamonds mark group means).

3.2. Personalised and On-Demand Mediation of Lexical Input

22 participants say GenAI provides abundant, user-demand mediation of vocabulary. These features—providing more than one context sentence, providing multiple register variations, providing audience-tailored explanations—were appreciated by learners as compared to traditional dictionaries. As P4 said, “If I type ‘arbitrary’ into ChatGPT, it will provide me with three sentences at three different levels – for a friend, for an email, for an essay; my dictionary can’t do this.” Likewise, P11 noted that: “I can ask it if it’s polite and what about the pragmatic part the AI gives?” Some were also “happy to be able to access” rapid Arabic–English contrastive glosses without having to switch between three apps: P7 commented, “It displays the English word, then explains it in Arabic when I request this - I didn’t need to switch between three apps.” Frequent usage of GenAI for vocabulary-related tasks included word meanings (n = 21) and translation (n = 20), “as seen in Figure 4”.

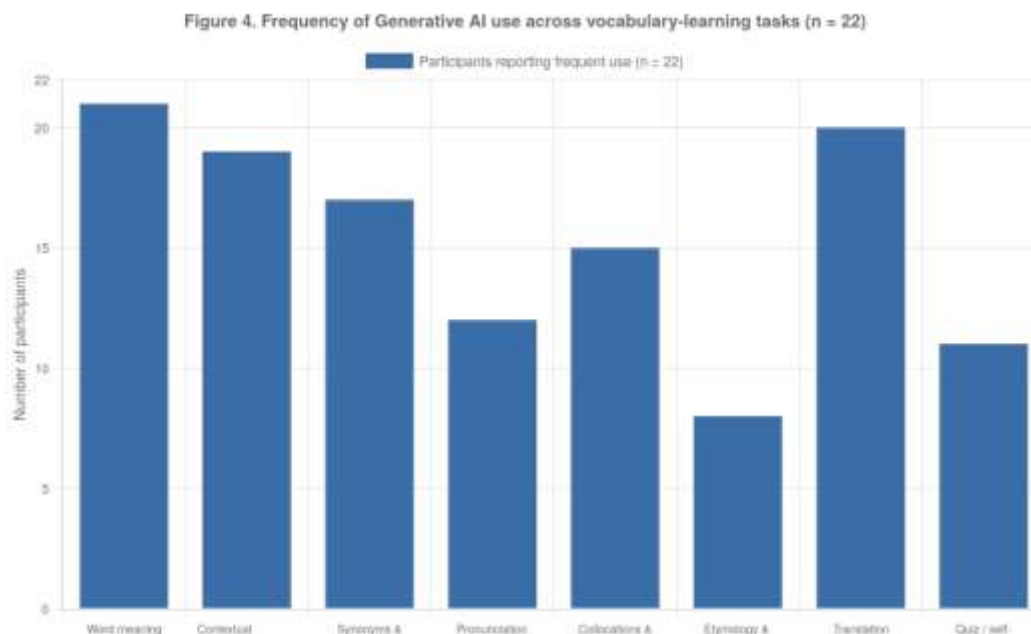


Figure 4. Frequency of Generative AI use across vocabulary-learning tasks (n = 22).

3.3. Affective Benefits: Confidence and Reduced Anxiety

19: GenAI as psychologically supportive. The non-judgemental (P8) and never bored (P17) aspects of the tool encouraged learners to ask questions at a primary/beginner level. P6, a B1 learner, said: “In class, I’m afraid to ask what ‘reckon’ means, with AI I can ask five times and no shame.” The learners identified the decreased anxiety with their tendency to try out lexically challenging performances, similar to previous studies in foreign-

language enjoyment and willingness to communicate [38,39]. Some evidence of the association between weekly use and self-efficacy ($\rho = .48$, $p = .023$) implies that greater practice of low-stakes quiz would be correlated with greater self-efficacy. Three participants, however, stated that this comfort became the avoidance of true human interaction which is explored here in more detail in section 3.5.

3.4. Prompt-Craft as a New Form of Learner Agency

A prominent inductive finding was learners' development of prompt-craft as a form of strategic competence. Seventeen participants described iterative prompt refinement: they began with simple queries ("Meaning of X") and progressed to elaborate instructions ("Give me five B2-level example sentences using 'arbitrary' in formal writing, and then a short paragraph on its collocations"). P14 said, "I am learning how to talk to it, and while I do that, I am also revising grammar and vocabulary." This is consistent with recent HAI scholarship suggesting that prompting is a bidirectional literacy practice [23,40]. Table 4 presents a taxonomy of learner prompt-types identified in the data. Quantitatively, the number of distinct prompt-types deployed rose sharply with CEFR level (B1 M = 2.71; C1 M = 6.40; $H(2) = 16.01$, $p < .001$, $d_{13} = 4.53$), and correlated strongly with weekly use ($\rho = .68$, $p < .001$) — the strongest single association observed. Figure 5 visualises the scatter between prompt-craft and self-efficacy, with a 95 % bootstrap band on the regression line.

Table 4. Taxonomy of learner prompt-types for vocabulary learning (N = 22).

Prompt type	Illustrative learner formulation	Mentions
Definition-seeking	"What is the meaning of 'ambivalent'?"	22
Contextualised examples	"Give me five sentences using 'ambivalent' in academic writing."	19
Level-tuned	"Explain at B1 level and then at C1 level."	13
Comparative / contrastive	"What is the difference between 'ambivalent' and 'indifferent'?"	15
Role-play / conversation	"Role-play a job interview and use these ten words."	11
Self-testing	"Give me a fill-in-the-blank quiz using today's words."	10
Meta-linguistic reflection	"Why did you choose this word here instead of another?"	7

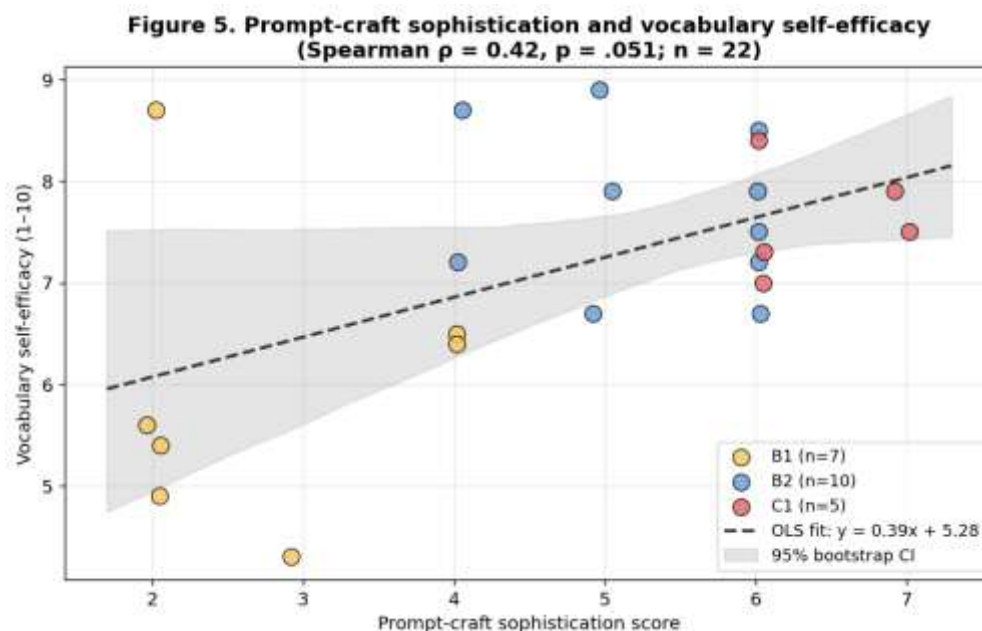


Figure 5. Prompt-craft sophistication and vocabulary self-efficacy, with OLS fit and 95 % bootstrap CI band; markers coloured by CEFR level.

3.5. Concerns and Tensions

Although excited about the opportunity, 20 participants expressed problems with accuracy, dependence and less human interaction. Errors were modest collocational infelicities to invented etymologies (“hallucinations”) [41]. P2 said, “It said that ‘arbitrary’ is from an Arabic word, which my teacher told me was wrong, it is Latin.” Nine said they used a dictionary or teachers to cross check their answers, developing what P19 described as a “reflex of doubt.” 10 participants issued concerns about relying entirely on AI, which they would not try to complete a task if they could not use their own brains. The fear of decreased interaction with human interlocutors was also an issue raised by 13 participants, particularly in the predominantly single-sex study contexts that some participants mentioned. These tensions are highlighted in Figure 6, which compares the number of people reporting benefits to the number reporting concerns. Partial support for this is found in the moderate positive correlation between benefit mentions and concern mentions ($\rho = .37$, $p = .088$), indicating that the critical awareness was not supplanted, but rather coexisted alongside, concerns about GenAI’s affordances.

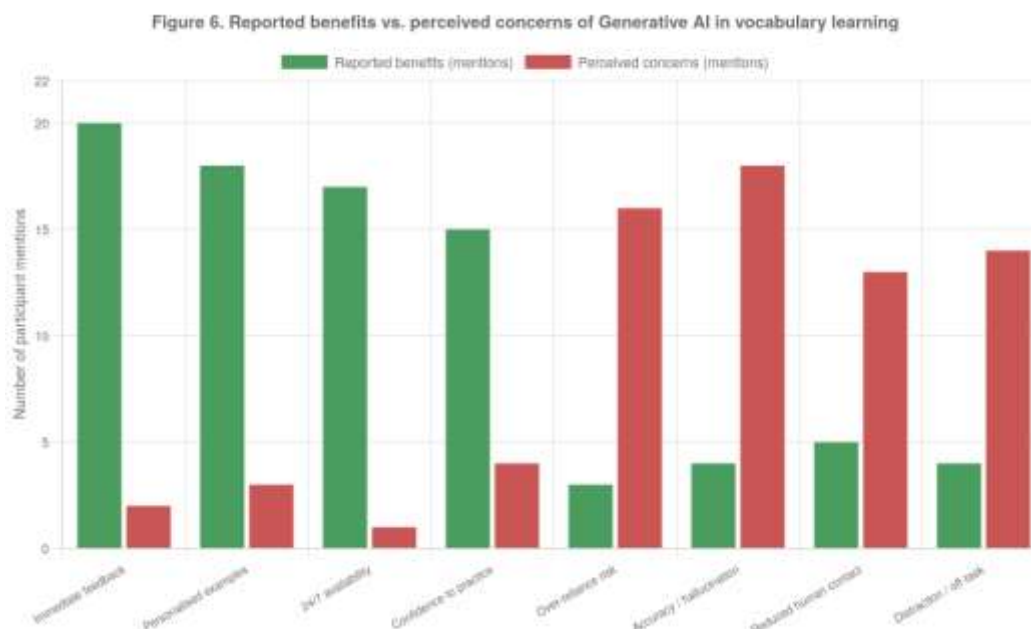


Figure 6. Reported benefits versus perceived concerns of Generative AI in vocabulary learning.

3.6. Reconfigured Learner–Teacher–Tool Ecology

The 18 participants placed GenAI in an expanded landscape of both “teachers and peers, as well as other resources”. Most participated in the discussion as agents of their own educators, seeing AI as a sort of “private tutor” (P3) or “second opinion” (P20) that would enhance their classroom teaching. P9 said, “In order for me to learn, I need my teacher, and I need the AI. For some, a new “sense of self-regulated learning and autonomy emerged”, aligning with current conceptions of learner autonomy in digital learning [42,43]. Five, however, were unsure why “teachers should be aware of or prohibit the use of AI”, marking a general discussion in academic integrity [44]. Specifically, “the chi-square test” of theme mentions against CEFR level ($\chi^2(8) = 0.59$, $p = .999$) found that there was no difference among the participants’ frequency of expressing a triadic conception (i.e., a learning environment consisting of “the learner, the instructor”, and the AI) across CEFR levels, meaning that this ecological framing was not exclusively limited to the more proficient ones.

3.7. Overview of Themes and Sub-themes

Table 5 summarises the five overarching themes, their constituent sub-themes, “and the number of participants” in whom each theme was coded. Prevalence “estimates should be interpreted alongside” the bootstrap confidence intervals reported in Figure 2.

Table 5. Overview of themes, sub-themes, and prevalence in the corpus (N = 22).

Theme	Sub-themes	Participants (n)
1. Personalised & on-demand mediation	Contextual examples; adaptive explanations; multiple synonyms; corpus-like retrieval	22

Theme	Sub-themes	Participants (n)
2. Affective benefits	Reduced anxiety; increased confidence; sustained motivation	19
3. Prompt-craft as new agency	Iterative refinement; role-play prompts; comparison across models	17
4. Concerns & tensions	Accuracy / hallucination; over-reliance; diminished human contact; distraction	20
5. Reconfigured learner–teacher–tool ecology	AI as tutor complement; changing role of teacher; identity as autonomous learner	18

4. Discussion

The findings extend existing scholarship on technology-mediated vocabulary learning by foregrounding the dialogic, adaptive, and affective dimensions of generative AI in a Saudi Arabian Arabic–English EFL context, and by triangulating qualitative accounts with quantitative descriptors. First, in line with earlier reviews of “CALL and mobile-assisted language learning” [10,45], learners experienced GenAI primarily as an extension of, rather than a replacement for, established tools. However, the depth of personalisation reported here surpasses what static dictionaries or curated apps can offer, echoing Kohnke et al.’s [14] observation that ChatGPT functions as a conversational partner capable of exemplifying pragmatic nuance. The “strong positive correlation between weekly frequency” and prompt sophistication ($\rho = .68$) suggests that engagement and strategic depth reinforce each other in a virtuous cycle, aligning with Nation’s [7] emphasis on multi-dimensional word knowledge and with corpus-linguistic accounts of collocational competence [46].

Second, the affective benefits reported by nineteen participants converge with “a growing body of work” on foreign-language enjoyment and anxiety [38,47]. The strong, though not overwhelming, correlation between weekly use and self-efficacy ($\rho = .48$, $p = .023$) suggests that low-risk, repeated, non-judgemental AI practice can enhance learner self-efficacy. This affordance could be particularly beneficial for learners at the lower-intermediate level who require repeated exposure, and should also draw on previous research on chatbots [48] but with much more sophisticated linguistic output. However, the discovery that three participants saw an elevated danger of the failure to interact authentically means “there is a need for pedagogical” focus. Remembering the prevailing role of human dialogue in L2 development in the sociocultural theory [21,22], teachers should make sure that tasks would be designed to provide input by both peers and teacher to ensure a balanced input.

Third, this study has added the concept of prompt-craft as a strategic competence as a unique contribution. There is also “a large effect size for the contrast between C1–B1” on prompt-craft ($d = 4.53$), indicating that prompt-craft could be used as an index of communicative development. Learners could not only ask questions, but also engage in metalinguistic dialog in an iterative fashion (which is considered to be a new type of language-learning strategy [8,49]). This is similar to recent HAI research that highlights “the co-constructed nature of the exchange” between human and LLM users [23,40] and Warschauer et al.’s [50] view of prompt engineering as a literacy. Appropriating a perspective of applied linguistics, prompt-craft brings lexical, syntactic, and pragmatic competence into focus, thus making the operation of teaching learners to prompt an act of language teaching.

Fourth, participants’ concerns regarding accuracy, over-reliance, and less human interaction resonate with the global conversation on the epistemic and social dangers of generative AI [16,17,44]. The positive correlation between benefits and concerns ($\rho = .37$) is interesting because it doesn’t seem to divide learners into enthusiastic and skeptical groups but seems to indicate that critical awareness and appreciation are co-developments. A growing ‘reflex of doubt’ (P19) is an encouraging sign of gaining AI literacy [51]. However, if not learned in a deliberate manner, these critical attitudes could be acquired and acquired unequally. Educational curricula should therefore embed opportunities for pupils to check against corpora, dictionaries and peer-reviewed sources, as well as opportunities for the pupils to consider ethical aspects of AI use [52,53]. Verification tasks can also be used to reveal subtle hallucinations in cross-linguistic etymological or morphological claims in Arabic–English contrastive setting such as that reported by P2.

To sum up, the reimagined learner–teacher–tool ecology found in this research is similar to the new models of hybrid or blended learning with AI [54,55]. The “chi-square test for the ecological framing” of AI as complement, rather than replacement, was not statistically significant across the CEFR levels ($\chi^2(8) = 0.59$, $p = .999$, $V = .09$)

showing that it was highly common across proficiency bands despite the significant differences in the behavioural indicators. In “another way of thinking, it wasn't a teacher vs. tool game”, rather it was a mutually dependent process, where the teacher gives direction, and AI executes practice, while the learner exercises judgment. This aligns with the sociocultural understandings of mediation and with recent calls for “AI-augmented pedagogy” [56]. It also emphasized the need to train teachers to build up their own specific literacies in the context of the Saudi and Gulf higher-education environments, because teachers have to develop these to guide students through “the affordances and limitations” of GenAI [57].

4.1. Limitations

There are several limits to this study. First, because of “the small sample size ($N = 22$)” and geographies, the statistical generalisations made are limited, and secondly, the patterns detected by the significant Kruskal–Wallis results must be viewed as patterns that the data could display, rather than as true patterns. Participant accounts are fairly detailed and the consistency of the L1–cultural background supports qualitative transferability. Secondly, as the study participants were all male, “results may not reflect the experience of female Saudi learners”, who may have different institutional and social contexts in which they acquire vocabulary using GenAI, and should be pursued in parallel. Finally, the fast-changing landscape of GenAI will result in different behaviours of tools reported in this document, but the central themes are likely to be the same.

5. Conclusion

This interview-based study, alongside non-parametric quantitative descriptions, sheds a light on “the ways in which Arabic-speaking” adult L2 learners of English experience generative AI in learning the L2 vocabulary in Saudi universities. The study is situated at the L2 acquisition and HAI nexus, empowering the learner's voice and supporting him/her in measured behavioural patterns. Processes of personalised, on-demand mediation of lexical input, affective benefits such as growth in confidence, and the new and emerging type of strategic competence, prompt-craft, were described by the learners. Data from quantitative descriptors proved that these behaviours increased in proportion to the level of attainment, and that strategic depth was enhanced by engagement, and critical awareness by appreciation.

Three implications follow. From a pedagogical perspective, teachers in Saudi and similar Arab EFL environments should structure activities that take advantage of the adaptive affordances offered by GenAI but also include tasks that rely on the human-agent aspect and call for critical verification. Prompts literacy and “AI literacy should be considered as complementary elements” in language programmes, curriculum-wise. Human–AI interaction should be highlighted in L2 vocabulary acquisition models, viewing the learning as an agentive interlocutor whose timely decisions contribute to the acquisition of L2 vocabulary. Future studies might attempt to validate the interview data using longitudinal interaction logs and use random comparisons of different instructional designs, as well as exploring other under-studied populations of learners, including female Saudi learners, and investigating the effect of Arabic–English contrastive prompting on the process of acquisition. Nevertheless, with the continued development of generative AI, continuous, qualitative, and quantitative focus on the learner's experience will continue to be vital to maintaining pedagogically relevant and responsible integration in “the field of second language” education and to upholding the inherent humanity of the technology.

References

1. Kasneci E, Sessler K, Kuchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ*. 2023;103:102274.
2. Bozkurt A, Xiao J, Lambert S, Pazurek A, Crompton H, Koseoglu S, et al. Speculative futures on ChatGPT and generative artificial intelligence: a collective reflection from the educational landscape. *Asian J Distance Educ*. 2023;18(1):53–130.
3. Zawacki-Richter O, Marin VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int J Educ Technol High Educ*. 2019;16:39.
4. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. *Educ Sci*. 2023;13(4):410.
5. Laufer B. The lexical plight in second language reading. In: Coady J, Huckin T, editors. *Second language vocabulary acquisition*. Cambridge: Cambridge University Press; 1997. p. 20–34.

6. Schmitt N, Schmitt D. Vocabulary in language teaching. 2nd ed. Cambridge: Cambridge University Press; 2020.
7. Nation ISP. Learning vocabulary in another language. 3rd ed. Cambridge: Cambridge University Press; 2022.
8. Schmitt N. Instructed second language vocabulary learning. *Lang Teach Res*. 2008;12(3):329–63.
9. Nakata T. Effects of expanding and equal spacing on second language vocabulary learning. *Stud Second Lang Acquis*. 2015;37(4):677–711.
10. Chen Z, Chen W, Jia J, An H. The effects of using mobile devices on language learning: a meta-analysis. *Educ Technol Res Dev*. 2020;68(4):1769–89.
11. Kohnke L. A pedagogical chatbot: a supplemental language learning tool. *RELC J*. 2023;54(3):828–38.
12. Hong W. The impact of ChatGPT on foreign language teaching and learning: opportunities in education and research. *J Educ Technol Innov*. 2023;5(1):37–45.
13. Yan D. Impact of ChatGPT on learners in an L2 writing practicum: an exploratory investigation. *Educ Inf Technol*. 2023;28(11):13943–67.
14. Kohnke L, Moorhouse BL, Zou D. ChatGPT for language teaching and learning. *RELC J*. 2023;54(2):537–50.
15. Baskara FXR. Chatbots and self-directed language learning: exploring learner engagement and outcomes. *J Engl Lang Pedagog Pract*. 2023;16(1):55–73.
16. Cotton DRE, Cotton PA, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. *Innov Educ Teach Int*. 2024;61(2):228–39.
17. Rudolph J, Tan S, Tan S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education? *J Appl Learn Teach*. 2023;6(1):342–63.
18. Godwin-Jones R. Distributed agency in second language learning and teaching through generative AI. *Lang Learn Technol*. 2024;28(1):5–31.
19. Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human–AI interaction. In: *Proc CHI Conf Hum Factors Comput Syst*. 2019. p. 1–13.
20. Shneiderman B. Human-centered AI. Oxford: Oxford University Press; 2022.
21. Lantolf JP, Poehner ME. Sociocultural theory and the pedagogical imperative in L2 education. New York: Routledge; 2014.
22. Vygotsky LS. Mind in society: the development of higher psychological processes. Cambridge (MA): Harvard University Press; 1978.
23. Jiang T, Sun Z, Fu S, Lv Y. Human–AI interaction research agenda: a user-centered perspective. *Data Inf Manag*. 2024;8(4):100078.
24. Alharbi W. AI in the foreign language classroom: a pedagogical overview of automated writing assistance tools. *Educ Res Int*. 2023;2023:4253331.
25. Song C, Song Y. Enhancing academic writing skills and motivation: assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Front Psychol*. 2023;14:1260843.
26. Mugableh AI. The impact of ChatGPT on the development of vocabulary knowledge of Saudi EFL students. *Arab World Engl J*. 2024;Special Issue on ChatGPT:265–81.
27. Dornyei Z, Ryan S. The psychology of the language learner revisited. New York: Routledge; 2015.
28. Tseng WT, Schmitt N. Toward a model of motivated vocabulary learning: a structural equation modeling approach. *Lang Learn*. 2008;58(2):357–400.
29. Bearman M, Ryan J, Ajjawi R. Discourses of artificial intelligence in higher education: a critical literature review. *High Educ*. 2023;86:369–85.
30. Reinders H, Benson P. Research agenda: language learning beyond the classroom. *Lang Teach*. 2017;50(4):561–78.
31. Lai C. Autonomous language learning with technology beyond the classroom. London: Bloomsbury Academic; 2017.
32. Kvale S, Brinkmann S. *InterViews: learning the craft of qualitative research interviewing*. 3rd ed. Los Angeles: Sage; 2015.
33. Temple B, Young A. Qualitative research and translation dilemmas. *Qual Res*. 2004;4(2):161–78.
34. Braun V, Clarke V. Thematic analysis: a practical guide. Los Angeles: Sage; 2022.
35. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–74.
36. Lincoln YS, Guba EG. *Naturalistic inquiry*. Newbury Park (CA): Sage; 1985.
37. Mackey A, Gass SM. *Second language research: methodology and design*. 3rd ed. New York: Routledge; 2022.

38. Dewaele JM, MacIntyre PD. Foreign language enjoyment and foreign language classroom anxiety: the right and left feet of the language learner. In: MacIntyre PD, Gregersen T, Mercer S, editors. *Positive psychology in SLA*. Bristol: Multilingual Matters; 2016. p. 215–36.
39. Peng JE, Woodrow L. Willingness to communicate in English: a model in the Chinese EFL classroom context. *Lang Learn*. 2010;60(4):834–76.
40. Zamfirescu-Pereira JD, Wong RY, Hartmann B, Yang Q. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In: *Proc CHI Conf Hum Factors Comput Syst*. 2023. p. 1–21.
41. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: *Proc ACM Conf Fairness Accountability Transparency*. 2021. p. 610–23.
42. Reinders H, White C. 20 years of autonomy and technology: how far have we come and where to next? *Lang Learn Technol*. 2016;20(2):143–54.
43. Benson P. Language learner autonomy: exploring teachers' perspectives on theory and practice. In: Chik A, Aoki N, Smith R, editors. *Autonomy in language learning and teaching*. London: Palgrave; 2018. p. 79–98.
44. Perkins M. Academic integrity considerations of AI large language models in the post-pandemic era: ChatGPT and beyond. *J Univ Teach Learn Pract*. 2023;20(2):7.
45. Golonka EM, Bowles AR, Frank VM, Richardson DL, Freynik S. Technologies for foreign language learning: a review of technology types and their effectiveness. *Comput Assist Lang Learn*. 2014;27(1):70–105.
46. McCarthy M, O'Keefe A. *Corpora and language teaching*. Cambridge: Cambridge University Press; 2010.
47. Botes E, Dewaele JM, Greiff S. The power to improve: effects of multilingualism and perceived proficiency on enjoyment and anxiety in foreign language learning. *Eur J Appl Linguist*. 2020;8(2):279–306.
48. Fryer LK, Ainley M, Thompson A, Gibson A, Sherlock Z. Stimulating and sustaining interest in a language course: an experimental comparison of chatbot and human task partners. *Comput Hum Behav*. 2017;75:461–68.
49. Oxford RL. *Teaching and researching language learning strategies: self-regulation in context*. 2nd ed. New York: Routledge; 2017.
50. Warschauer M, Tseng W, Yim S, Webster T, Jacob S, Du Q, et al. The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *J Second Lang Writ*. 2023;62:101071.
51. Ng DTK, Leung JKL, Chu SKW, Qiao MS. Conceptualizing AI literacy: an exploratory review. *Comput Educ Artif Intell*. 2021;2:100041.
52. Boulton A, Cobb T. Corpus use in language learning: a meta-analysis. *Lang Learn*. 2017;67(2):348–93.
53. Selwyn N. On the limits of artificial intelligence (AI) in education. *Nord J Stud Educ Policy*. 2024;10(1):3–14.
54. Pack A, Maloney J. Using artificial intelligence in TESOL: some ethical and pedagogical considerations. *TESOL Q*. 2024;58(2):1007–18.
55. Ranalli J. L2 student engagement with automated feedback on writing: potential for learning and issues of trust. *J Second Lang Writ*. 2021;52:100816.
56. Holmes W, Tuomi I. State of the art and practice in AI in education. *Eur J Educ*. 2022;57(4):542–70.
57. Moorhouse BL, Kohnke L. The effects of generative AI on initial language teacher education: the perceptions of teacher educators. *System*. 2024;122:103290.