

Search in Transition: A Mixed-Methods Study of LLMs and Search Engines in English Test Preparation

Tarek Rahman

United International University

Md Shaharia Hossen

`mhossen201288@bscse.uiu.ac.bd`

United International University

Mark Protik Mondol

United International University

Jannatun Noor Mukta

United International University

Article

Keywords: Large Language Models in Education, English Language Test Preparation, Hybrid Search and Learning Environments, Human–Computer Interaction, AI-Supported Learning

Posted Date: October 21st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10313360/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Search in Transition: A Mixed-Methods Study of LLMs and Search Engines in English Test Preparation

Abstract

As large language models (LLMs) increasingly shape academic information-seeking, university students are adopting hybrid workflows that combine conversational AI and traditional search engines. Yet little is known about how this integration influences interactive learning in high-stakes contexts such as English language test preparation. Grounded in cognitive load theory and self-regulated learning perspectives, this mixed-methods study examines how students strategically use LLM-based tools and search engines during English test problem-solving for higher studies. Survey data from 140 students were analysed using repeated-measures ANOVA and chi-square tests, complemented by task-based interviews with 20 participants completing seven representative test activities. Findings show that LLMs are perceived as more efficient and easier to use, whereas search engines are preferred for verification and rule confirmation. Students employing a balanced hybrid strategy achieved the highest task accuracy despite longer completion times, indicating a trade-off between efficiency and reliability. Qualitative results suggest that tool selection is context-dependent and shaped by learners' efforts to manage cognitive load and maintain epistemic trust. The study proposes a hybrid interface model that embeds LLM assistance within search environments to support trustworthy, AI-mediated interactive learning in higher education.

Keywords: Large Language Models in Education; English Language Test Preparation; Hybrid Search and Learning Environments; Human–Computer Interaction; AI-Supported Learning

1 Introduction

Since the advent of Artificial Intelligence (AI) and Large Language Models (LLMs)⁹, the way that university students handle information, read textual content, and anticipate examinations has changed drastically. Comprehensiveness and authoritative search are key features of the Internet, which have enabled students to learn grammar rules, words from the Internet, and pick up sample or official exam questions. Conversational AI models such as ChatGPT of OpenAI have recently contributed a novel paradigm of including explanations, linguistic aids, and summaries in a natural way using a conversation, which could be adopted as an efficient way to learn English and prepare for English language proficiency examinations¹.

As students' use in this sphere is growing, it is important to understand how students are using and how students are judging the various tools they are using to get information for high-stakes assessments in English. Previous studies have found that LLMs are best trained for language-related tasks such as summarizing, paraphrasing, explanations, and response writing for passing examinations such as IELTS, TOEFL, etc in a limited time frame, which was one of the most important requirements for passing an examination. Yet, they may vary in performance when it comes to various problems and different prompts because of the linguistic attributes of the problem at hand^{5,13}. Unlike classical search engines, they can assess official resources and grammar resources, look for input from several opinions, but are also bound to do so and to navigate through a multitude of web pages in search of reliable information. Hence, more recent studies appear to be moving towards considering LLMs and search engines as complementary tools – using the conversational AI to provide explanations and generate content, but search engines to verify facts or language rules generated by the AI (for example, by the AI^{4,11}).

Either used alone is satisfactory but does not address all of the needs towards successful IELTS test preparation. Some answers generated by LLMs may also be fluent and coherent, but false answers that don't contain contextually relevant information for the query may negatively impact the learner if fed back to them, with no modifications¹³. Traditional search engines, on the other hand, can be used to find credible and verifiable sources; however, it still takes a lot of work to scour sources, compare, and synthesize the information in multiple web pages. This can lead to a loss in efficiency due to the normal process of thinking and working, and can add to the difficulty of the task – especially if there is a time limit such as occurs with a standardized test. As a result, many students tend to adopt hybrid search techniques where using the tools with LLMs along with the search results provided by the traditional search engine to make use of the benefits of both, as they are both efficient and offer high-quality and trustworthy information in the search results^{10,6}. Reusing several platforms over time, however, could disrupt users' activities, put more load on the user's cognitive capacities, and can have a negative impact on interactive learning in complex problem-solving situations.

Although there are a growing number of studies on how LLMs and SEs can be used in the field of education and information retrieval, there is a lack of research concerning the strategic use of LLMs and SEs in preparing for ELTs. Research so far has focused either on one or the other of these tools and/or on a single measure of performance, but has, to date, produced few answers to the questions of how the learners decide how to choose a strategy, what this strategy is, and what conceptions about their work emerge. Secondly, very little is known about the tradeoffs that students make between efficiency and usability on the one hand and 'trust' and 'accuracy' on the other hand when using conversational AI and 'old school' search engaged in language learning activities.

This study is intended to explore LLM-based tools' attitudes and usage, as well as the preference of university students towards these tools compared to the traditional search engine and the combination of both tools, in preparing for undergraduate language exams in the field of English. In this study, researchers apply a Mixed Methods Research Design whereby quantitative analysis evidence is gathered from the results of survey responses from 140 University students, and the qualitative analysis evidence is gathered from the results of interviews based on a task from 20 University students. The two in tandem can be used to assess the different dimensions of use, use efficiency, satisfaction, user workflow, user expectations, the good and bad of conversational AI and information retrieval.

Specifically, with respect to research, this study will try to address the following questions:

- **RQ1:** How do university students use LLM-based tools and traditional search engines for English language test preparation?
- **RQ2:** Are there significant differences in perceived usage frequency, satisfaction, efficiency, and ease of use between LLM-based tools and traditional search engines?
- **RQ3:** How do different tool usage strategies affect task accuracy and completion time during English language test problem solving?
- **RQ4:** What factors influence students' preferences for LLMs, search engines, or hybrid tool usage?

Based on the mixed-methods investigation, this research makes contributions in the following ways:

1. It provides real evidence to assess and compare the performances of LLM-based tools and the traditional search engine tools in the English exam preparation task of IELTS and TOEFL.
2. It provides accuracy data for information posts, providing a numeric value of proficiency and error on performing the task based on the use of various information seeking strategies.
3. It features qualitative, descriptive analytics of learners' behaviours, cognition, and trust with transitions from conversational AI to traditional search. With the main advantages of using LLMs for learner assistance being greater efficiency in information retrieval, efficiency in writing, and better user experience, it proposes a human-computer interaction (HCI)-oriented hybrid interaction model by mitigating the potential risks and streamlining the interactions for the LLM.

This work was motivated by the preprint version of this paper, which is at ¹⁵. The present manuscript represents a significant improvement over this preliminary version with better methodological reporting, increased statistical and qualitative analyses, a better discussion, documentation for reproducibility, and edits for transparency and research reporting.

2 Literature Review

Indeed, the emerging world of Large Language Models (LLMs) has increasingly influenced students' approach to language-related tasks like the incorporation of L2 in the L2 language skills of university students taking EELTs. While there is extensive research on the usability and trust of LLMs in recent years, and the possibilities of using them to prepare for language exams, few studies have focused on comparing their performance in providing grammar explanation rules and vocabulary explanations to that of traditional search engines, such as Google. Divekar et al.⁵ studied the impact of which model students prefer when learning (LLMs or traditional search engines), investigated the strengths and weaknesses of LLMs in the context of the properties of the tasks, and reached the conclusion that it works well on more straightforward tasks, though it will not work on more complex tasks. In the same vein, Kumar et al.⁷ showed that the LLM could help to better structure language tasks and enhance student learning for language tasks such as writing and paraphrasing.

There are several studies that have placed an emphasis on an individual's achievement of goals and performance effectiveness. Spatharioti et al.¹¹ came up with a randomized experiment to prove that users of LLMs are more efficient in decision-making and require fewer queries. But something they are not able to capture, which is a problem with users – even when all answers are incorrect, there's a lack of convincingness in the correctness they provide, but a lot of fluidity and confidence. This worry was further enhanced by Xu et al.¹³ and their emphasis on the need for appropriate verification to provide when making use of content generated by LLMs. This could affect factual information and grammar if the test-taker is an ELL who takes an English test without referring to the model output.

Task preference studies also suggest that the complementary properties of LLMs and traditional search engines can exist. These scenarios have been explored by Caramancion⁴: they found that users indicated that using conversational ChatGPT-style interfaces when seeking information led to a level of personalisation, efficiency, and convenience for the users. The results of these findings were verified by Sakirin and Said¹⁰. On top of this, Wazzan et al.¹² have shown that users' strategy is affected by the use of tools, too: while the interaction with the LLMs is much easier to understand, users opting for the classical tools must follow a structured navigation – and this is what is done when preparing for periodic English language assessments.

There continue to be issues of credibility and transparency of sources. While AI tools may provide convenience and speed, for academic and educational research authoritative sources are required; hence traditional search is recommended as follows by Kapoor et al.⁶. The LLMs do not have a reference pointed out, however, which could create issues for learners who might want to have the rules 'checked' in the SLA classroom. This compromise has suggested that the systems of mixed plant & animal may be used. Previous research focused on metasearch engines to address issues of aggregating information from multiple sources to achieve better accuracy, while more recent studies were very compelling concerning using contextual reasoning of LLMs and the ability to find verifications from search engines^{4,3}.

Although there is increasing attention on hybrid systems, the previous reviews appear to have solely investigated LLMs or classic search engines or only investigated them within certain tightly framed tasks, and don't fully reflect how students synergetically use these two systems when preparing for English language tests¹³. There are a lot of efforts filling this void, such as Perplexity AI, which aims to add a link from the AI's answer to the websites that the AI used as information sources⁸, or Google that's built AI-generated summaries in Google Search in its latest experience, Search Generative Experience (SGE)⁸. Although the tool will, to some degree, try to resolve this issue by offering links to the content used to generate the answer by the AI⁸, and Google implemented AI-generated summaries in the Search experience with the new Search Generative Experience (SGE), there still lacks personalized, adaptive, and task-specific pedagogical support. This current study attempts to fill these gaps by taking a mixed-methods approach to investigating the students' hybrid tool use behaviors when solving problems in the language test domain and by designing a prototype based on the users' behaviors and contexts in the domain.

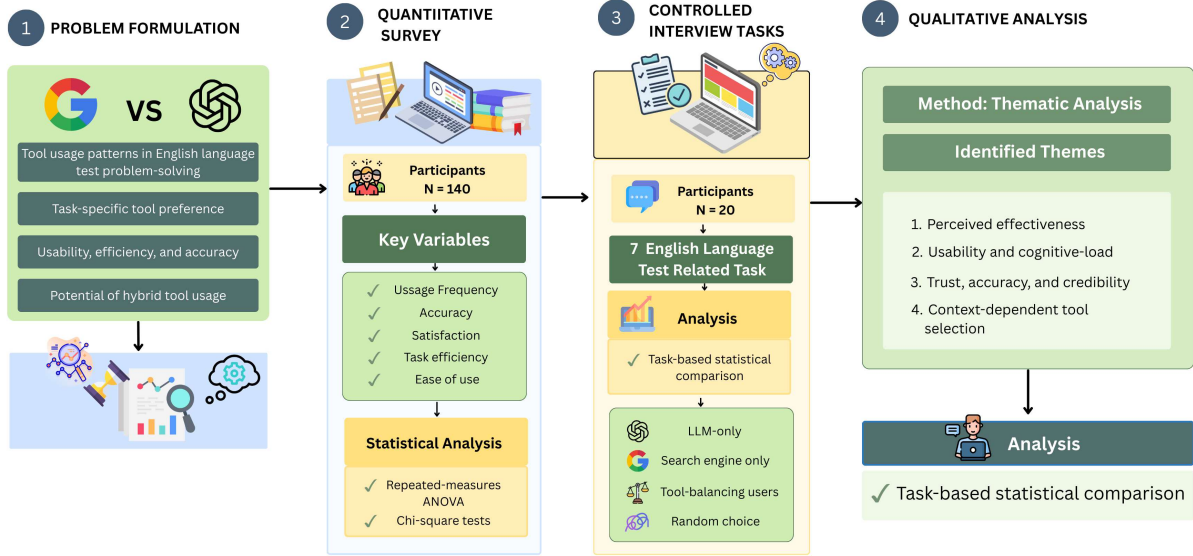


Figure 1: Overview of the study methodology. A mixed-methods approach was employed in this study. (1) The survey phase ($n = 140$) captured quantitative data and was analyzed using different statistical tests. (2) The qualitative phase included in-person interviews ($n = 20$), where participants completed seven English language test problem tasks and were grouped based on tool usage. Thematic analysis of open-ended responses and interview transcripts led to four themes.

3 Methodology

To explore university students’ preferences and usage behaviors regarding large language models (LLMs) and traditional search engines in the context of English language test problem solving for higher studies, we employed a mixed-methods approach that combined quantitative and qualitative data collection and analysis, as illustrated in Figure 1. This design enabled the examination of both broad usage patterns and deeper user experiences in a complementary manner.

We first conducted an online survey that collected responses from 140 university students across a range of academic disciplines. While the survey primarily targeted students from technology-related fields such as Computer Science and Engineering (CSE), Electrical and Electronics Engineering (EEE), and Data Science, it was also distributed to students from other disciplines, including Business Administration and English studies, to ensure diversity. The questionnaire included both closed-ended and open-ended questions designed to assess tool usage frequency, satisfaction, efficiency, and perceived ease of use when employing GPT-based LLMs and traditional search engines such as Google for English language test preparation. Descriptive statistics were used to summarize the data, and inferential statistical analyses, including one-way ANOVA and chi-square tests, were conducted to evaluate differences in user perceptions and the influence of demographic variables on tool preference.

To enrich and validate the survey findings, we conducted in-depth, in-person interviews with 20 university students from CSE, EEE, BBA, and English backgrounds, primarily recruited from United International University (UIU), Dhaka, Bangladesh. Each participant was asked to complete seven English language test-related tasks representative of higher studies preparation. These tasks included summarizing an academic passage, correcting grammatical errors, selecting appropriate vocabulary in context, paraphrasing sentences, drafting a short opinion-based response similar to IELTS Writing Task 2, drafting a response similar to IELTS Writing Task 1, and interpreting information from a short academic text.

Participants were categorized into four groups based on their tool usage behavior during task completion: LLM-only users, search engine-only users, balanced users who alternated between both tools, and free-choice users who switched between LLMs and search engines based on task demands.

The qualitative analysis integrated insights from both open-ended survey responses and interview transcripts. A thematic coding approach was applied, focusing on perceived effectiveness, usability, trustworthiness, and contextual factors influencing tool selection. This analysis offered a comprehensive understanding of how students navigate the strengths and limitations of LLMs and traditional search engines in English language test problem solving, and how their choices are shaped by task type, familiarity with English language rules, and perceived cognitive effort.

3.1 Survey

The survey included **140 university students** from diverse academic disciplines and demographic backgrounds. Although Computer Science and Engineering (CSE) students constituted the largest group, the sample also included participants from Electrical and Electronics Engineering (EEE), Business Administration (BBA), Data Science, English, and other academic disciplines. Table 1 summarizes the distribution of participants by department, CGPA range, gender, and age group.

The survey instrument was designed to examine students’ use and perceptions of traditional search engines (e.g., Google) and Large Language Model (LLM)-based tools (e.g., ChatGPT) for English language test preparation in the context of higher studies, including tests such as IELTS and TOEFL. The questionnaire included demographic questions, tool-usage questions, comparative perception measures, an overall tool-preference question, and open-ended questions concerning task-specific preferences and perceived limitations.

The quantitative items assessed four core dimensions for both traditional search engines and LLM-based tools:

- frequency of use for English test preparation;
- satisfaction with the accuracy of information and solutions provided;
- perceived efficiency in completing English test tasks; and
- ease of use for English test problem solving.

Participants evaluated LLM-based tools and traditional search engines separately. The response scales were tailored to the construct being measured. Usage frequency was assessed using ordered frequency categories ranging from “Never” to “Always.” Satisfaction was measured using a five-point scale ranging from “Very Dissatisfied” to “Very Satisfied.” Efficiency was assessed using a scale ranging from “Not Efficient at All” to “Extremely Efficient,” while ease of use was measured using a scale ranging from “Very Difficult” to “Very Easy.”

At the end of the survey, participants were asked to indicate their overall preference among LLM-based tools, traditional search engines, or a hybrid strategy involving both. The questionnaire also included open-ended questions asking participants to identify the English test tasks for which they preferred LLMs over search engines and to describe limitations or challenges encountered when using either type of tool.

This parallel evaluation structure enabled systematic comparisons between LLM-based tools and traditional search engines across the four core dimensions. At the same time, the overall-preference and open-ended questions provided additional insight into students’ broader tool-selection strategies and task-specific experiences in English test preparation. Overall summary of the Survey Statistics in 1.

3.2 In-person Interview

To complement the survey findings and gain deeper insights into students’ tool usage behaviors, we conducted in-person interviews with **20 university students** from diverse academic backgrounds, including Computer Science and Engineering (CSE), Electrical and Electronics Engineering (EEE), Business Administration (BBA), and English studies backgrounds, primarily recruited from United International University (UIU), Dhaka, Bangladesh.

The interview protocol consisted of a structured sequence of **seven English language test-related tasks** designed to reflect common activities involved in preparation for higher studies. Participants completed seven English language test-related tasks: (1) summarizing an English language test based, (2)

Department	
Computer Science & Engineering (CSE)	72
Electrical & Electronics Engineering (EEE)	24
Business Administration (BBA)	18
Data Science	12
English	10
Other	4
CGPA Range	
3.81–4.00	25
3.51–3.80	29
3.01–3.50	62
2.50–3.00	24
Gender	
Male	93
Female	47
Age Range	
18–20 years	9
21–25 years	128
26–30 years	3

Table 1: Distribution of participants by department, CGPA range, gender, and age group

correcting grammatical errors in a language testing problem, (3) selecting appropriate vocabulary in context, (4) paraphrasing sentences without altering meaning, (5) drafting an opinion-based response similar to IELTS Writing Task 2, (6) drafting a data description response similar to IELTS Writing Task 1, and (7) interpreting information from an academic reading passage. The tasks were selected through consultation with English language instructors and a review of common IELTS and TOEFL preparation materials to ensure contextual relevance and varying cognitive demands. The primary objective was to observe how tool choice influenced task strategy, accuracy, and efficiency across different English language test activities.

To assess task performance, we developed task-specific evaluation rubrics in consultation with faculty members specializing in English language teaching. For instance, summarization tasks were evaluated based on coherence, coverage, and conciseness; grammar and vocabulary tasks were assessed based on correctness and appropriateness; and writing tasks were scored according to clarity, coherence, and task fulfillment. Each task was independently evaluated by two raters to ensure inter-rater reliability.

Participants were categorized into four groups based on their tool usage behavior during task completion: (1) **LLM-only users**, (2) **search engine-only users**, (3) **tool-balancing users** who sequentially used both LLMs and search engines, and (4) **free-choice users** who selected tools flexibly depending on task demands. This grouping enabled comparisons of accuracy and task completion time across different task types and facilitated analysis of how tool-switching behavior related to user preferences and task complexity.

In addition to task performance, the interviews included open-ended reflections on perceived usability, trustworthiness, and strengths or limitations of each tool. These qualitative responses were analyzed using thematic coding and were used to contextualize quantitative findings and inform the design considerations for a hybrid English language test support system.

3.3 Research Hypotheses

The following hypotheses were evaluated:

- H_0 : There is no statistically significant difference between LLM-based tools and traditional search engines.
- H_1 : There is a statistically significant difference between LLM-based tools and traditional search engines.

3.4 Statistical Analysis

The quantitative survey data were analyzed using descriptive and inferential statistical techniques. All analyses were performed using Python 3.11 with the Pandas, NumPy, SciPy, and Matplotlib libraries. Responses were first coded numerically using five-point ordered scales corresponding to each construct. Usage frequency responses were coded from Never (1) to Always (5). Satisfaction responses were coded from Very Dissatisfied (1) to Very Satisfied (5). Efficiency responses were coded from Not Efficient at All (1) to Extremely Efficient (5), while ease-of-use responses were coded from Very Difficult (1) to Very Easy (5).

Descriptive statistics were computed to summarize participant responses for each quantitative variable. Inferential statistical analyses were subsequently conducted to compare participants' evaluations of LLM-based tools and traditional search engines. Because every participant evaluated both tool categories, paired non-parametric statistical tests were selected. Statistical significance was evaluated using a significance level of $\alpha = 0.05$.

3.4.1 Descriptive Statistics

Descriptive statistics were calculated to summarize the distribution of each quantitative variable. The reported statistics included the sample size (N), mean ($\bar{x}$), standard deviation (s), median, first quartile (Q_1), third quartile (Q_3), interquartile range (IQR), minimum, maximum, and 95% confidence interval.

The arithmetic mean was computed as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

where x_i denotes the observed value of participant i , and n is the total number of observations.

The sample standard deviation was calculated using

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}. \quad (2)$$

The 95% confidence interval of the mean was estimated as

$$CI = \bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \quad (3)$$

where $t_{\alpha/2, n-1}$ denotes the critical value of the Student's t distribution.

3.4.2 Wilcoxon Signed-Rank Test

Since the same participants evaluated both LLM-based tools and traditional search engines, paired comparisons were performed using the two-sided Wilcoxon signed-rank test. This non-parametric test is appropriate for paired ordinal data obtained from Likert-scale responses and does not require the assumption of normality.

The Wilcoxon test statistic is defined as

$$W = \min(W^+, W^-), \quad (4)$$

where W^+ and W^- represent the sums of the positive and negative signed ranks, respectively.

To quantify the magnitude of paired differences, the rank-biserial correlation was computed as

$$r_{rb} = \frac{W^+ - W^-}{W^+ + W^-}. \quad (5)$$

The Wilcoxon signed-rank test was applied separately to compare usage frequency, satisfaction, efficiency, and ease-of-use ratings between LLM-based tools and traditional search engines.

3.4.3 Friedman Test

To determine whether participants rated the four measured dimensions (usage frequency, satisfaction, efficiency, and ease of use) differently within each tool category, the Friedman test was conducted. The Friedman test is a non-parametric alternative to repeated-measures ANOVA and is suitable for repeated ordinal measurements.

The Friedman test statistic is given by

$$\chi_F^2 = \frac{12}{nk(k+1)} \sum_{j=1}^k R_j^2 - 3n(k+1), \quad (6)$$

where n denotes the number of participants, k represents the number of repeated measurements, and R_j is the sum of ranks for the j^{th} measurement.

The corresponding effect size was estimated using Kendall’s coefficient of concordance:

$$W = \frac{\chi_F^2}{n(k-1)}. \quad (7)$$

3.4.4 Chi-square Test of Independence

A chi-square test of independence was conducted to examine whether participants’ gender was associated with their overall tool preference.

The chi-square statistic was computed as

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \quad (8)$$

where O_i and E_i denote the observed and expected frequencies, respectively.

The strength of the association was measured using Cramer’s V :

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}, \quad (9)$$

where n is the sample size and k is the smaller number of rows or columns in the contingency table.

The chi-square analysis was used to investigate the relationship between gender and participants’ overall preference for LLM-based tools, traditional search engines, or a hybrid approach.

3.5 Ethical Considerations

Participation in the study was voluntary. All participants provided informed consent before completing the survey and interview tasks. Personally identifiable information was removed prior to analysis to ensure participant confidentiality. The collected data were used exclusively for academic research purposes.

4 Quantitative Analysis

This section summarizes the quantitative findings of the survey. It presents descriptive statistics and comparative analyses of participants’ demographics, tool preferences, usage frequency, satisfaction, perceived efficiency, and ease of use for LLMs and search engines in English test preparation.

4.1 Survey Results

A total of 140 valid survey responses were included in the quantitative analysis. The quantitative responses were coded using five-point ordered response scales according to the construct being measured. For usage frequency, responses were coded as Never = 1, Occasionally = 2, Sometimes = 3, Often = 4, and Always = 5. Satisfaction responses were coded from Very Dissatisfied = 1 to Very Satisfied = 5. Efficiency responses

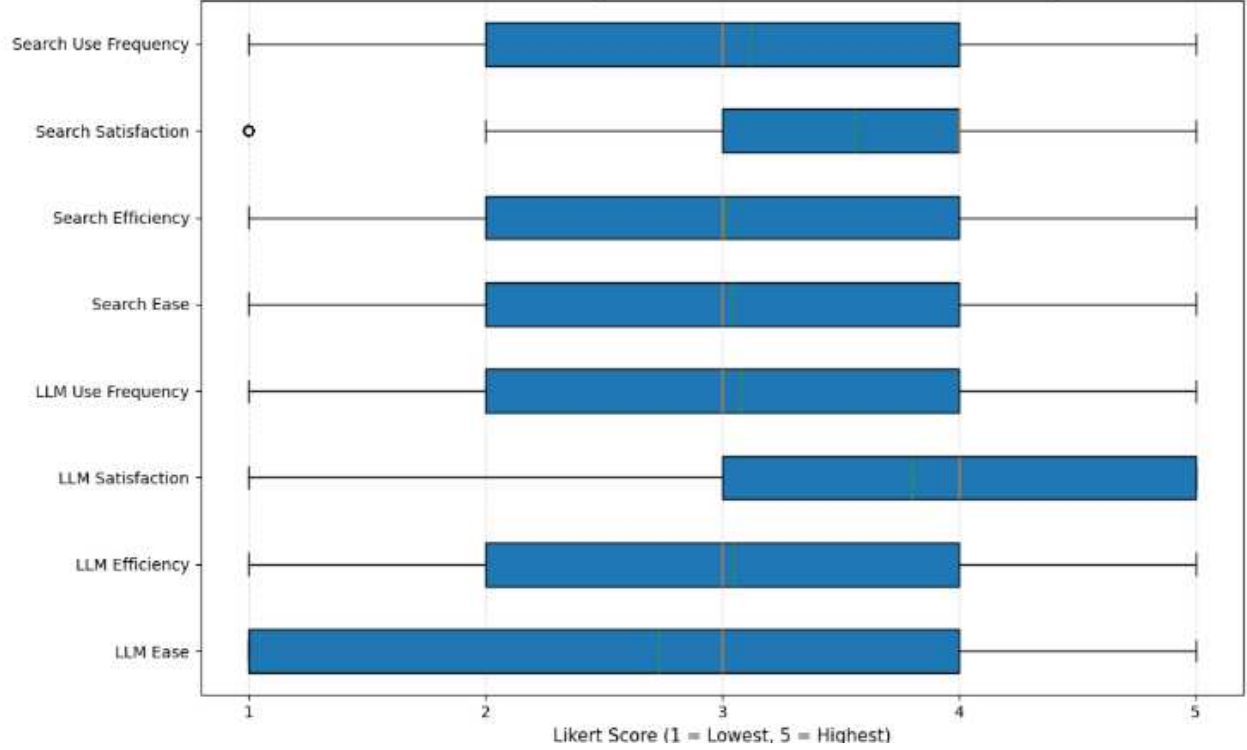


Figure 2: Distribution of quantitative survey measures for LLM-based tools and traditional search engines. The figure presents usage frequency, satisfaction, efficiency, and ease-of-use ratings on five-point scales.

were coded from Not Efficient at All = 1 to Extremely Efficient = 5. Ease-of-use responses were coded from Very Difficult = 1 to Very Easy = 5.

Descriptive statistics were calculated for four dimensions of tool use: usage frequency, satisfaction, efficiency, and ease of use. These measures were assessed separately for LLM-based tools and traditional search engines.

The descriptive results showed that the mean usage-frequency score was slightly higher for traditional search engines ($M = 3.121$, $SD = 1.349$) than for LLM-based tools ($M = 3.079$, $SD = 1.469$). In contrast, LLM-based tools received a higher mean satisfaction score ($M = 3.800$, $SD = 1.114$) than traditional search engines ($M = 3.557$, $SD = 1.081$). The mean efficiency scores were similar, with LLM-based tools receiving a slightly higher score ($M = 3.050$, $SD = 1.380$) than traditional search engines ($M = 3.014$, $SD = 1.275$). For ease of use, traditional search engines received a higher mean score ($M = 3.043$, $SD = 1.424$) than LLM-based tools ($M = 2.736$, $SD = 1.442$).

These descriptive results suggest that the two tool categories demonstrated different patterns across the measured dimensions rather than a uniformly superior performance by one tool category. The distribution of responses across the eight quantitative measures is illustrated in Figure 2.

4.2 In-Person Interview

To analyze the data collected during the in-person interview phase, both quantitative and qualitative aspects of participant performance were examined while participants completed a series of structured English language test-related tasks. A total of **20 students** participated in this phase of the study. Each participant completed seven tasks representing common English language test preparation activities and covering a range of cognitive demands, including comprehension, grammatical accuracy, paraphrasing, and written expression.

Participants were grouped according to their tool-usage strategy during task completion: **LLM-only users**, **search-engine-only users**, **tool-balancing users** who used both tools sequentially, and **flexible-**

Tool-Usage Group	Accuracy (%)	Time (min)
LLM-only	88	24
Search-engine-only	80	31
Tool-balancing	90	35
Flexible-choice	82–88	25–32

Table 2: Summary of task performance by tool-usage strategy in the in-person interview phase.

choice users who selected tools according to task requirements. Task performance was evaluated using a predefined rubric, and completion time was recorded for each participant. Table 2 summarizes the reported quantitative findings from the in-person interview phase.

4.2.1 Accuracy Analysis

Each participant’s response was evaluated using a predefined scoring rubric tailored to the specific English language test task. Summarization tasks were assessed based on coverage, conciseness, and coherence; grammar and vocabulary tasks were evaluated for correctness and appropriateness; and writing tasks were assessed based on clarity, organization, and task fulfillment. The use of a predefined rubric was intended to promote consistent evaluation across the assessed tasks.

Participants who relied exclusively on LLM-based tools achieved an average accuracy of **88%**. Participants who relied exclusively on traditional search engines achieved an average accuracy of **80%**. The tool-balancing group, which combined LLMs and search engines, achieved the highest reported average accuracy of **90%**. Participants in the flexible-choice group achieved accuracy scores ranging from **82% to 88%**.

These findings suggest that the combination of LLM-based tools and traditional search engines may provide complementary advantages for different English language test tasks. However, because the interview phase involved a relatively small sample of 20 participants and the reported group-level results are descriptive, these findings should be interpreted as exploratory rather than as evidence of statistically significant differences between tool-usage strategies.

4.2.2 Completion Time Analysis

Completion time was also examined across the different tool-usage strategies. Participants in the LLM-only group completed the assigned tasks in an average of approximately **24 minutes**, whereas the search-engine-only group required approximately **31 minutes**. The tool-balancing group had the longest reported average completion time, approximately **35 minutes**. The flexible-choice group demonstrated completion times ranging from approximately **25 to 32 minutes**.

The descriptive results suggest a potential trade-off between task completion time and accuracy. The LLM-only strategy was associated with the shortest reported completion time, whereas the tool-balancing strategy was associated with the highest reported accuracy and the longest reported completion time. These observations are descriptive and should not be interpreted as evidence of a causal relationship between tool-usage strategy, accuracy, and completion time.

Figure 3 illustrates the descriptive relationship between accuracy and completion time across the reported tool-usage strategies.

4.2.3 Inter-Rater Reliability

To assess consistency in task evaluation, two independent raters evaluated participant responses using the predefined scoring rubrics. Inter-rater reliability was assessed using Cohen’s kappa. The reported Cohen’s kappa value was $\kappa = 0.82$, indicating a high level of agreement between the two raters. Any discrepancies were resolved through discussion and consensus.

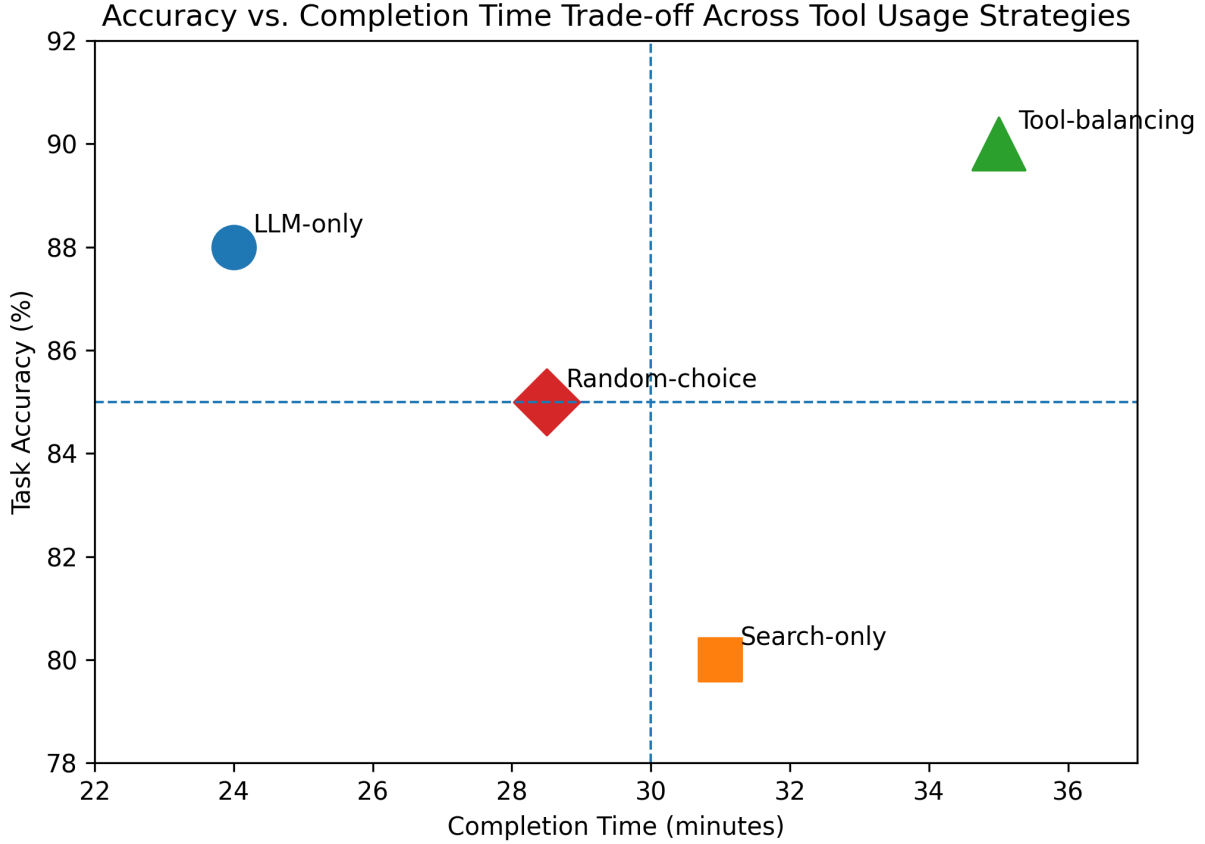


Figure 3: Descriptive relationship between task accuracy and completion time across the different tool-usage strategies in English language test problem solving.

4.3 Paired Comparison Between LLM-Based Tools and Search Engines

Because the same participants evaluated both LLM-based tools and traditional search engines, paired comparisons were conducted using the Wilcoxon signed-rank test. This non-parametric test was selected because the responses were collected using ordered Likert-scale categories. Figure 4 visually summarizes these comparisons together with the corresponding confidence intervals and statistical test results.

The Wilcoxon signed-rank tests did not identify statistically significant differences between LLM-based tools and traditional search engines for any of the four evaluated dimensions at the $\alpha = 0.05$ significance level. Usage frequency did not differ significantly between LLM-based tools and traditional search engines ($W = 3352.5$, $p = .784$, rank-biserial $r = -.029$). Satisfaction showed a non-significant trend favoring LLM-based tools ($W = 2230.0$, $p = .070$, rank-biserial $r = .199$). The difference in perceived efficiency was also not statistically significant ($W = 2960.5$, $p = .659$, rank-biserial $r = .048$). Similarly, the difference in ease of use was not statistically significant ($W = 2683.0$, $p = .066$, rank-biserial $r = -.196$).

Although the descriptive mean was higher for LLM-based tools in satisfaction and efficiency, these differences did not reach conventional statistical significance. Conversely, traditional search engines had higher descriptive mean scores for usage frequency and ease of use, but these differences were also not statistically significant.

4.4 Within-Tool Differences

To examine whether participants' ratings differed across the four measured dimensions within each tool category, separate Friedman tests were conducted.

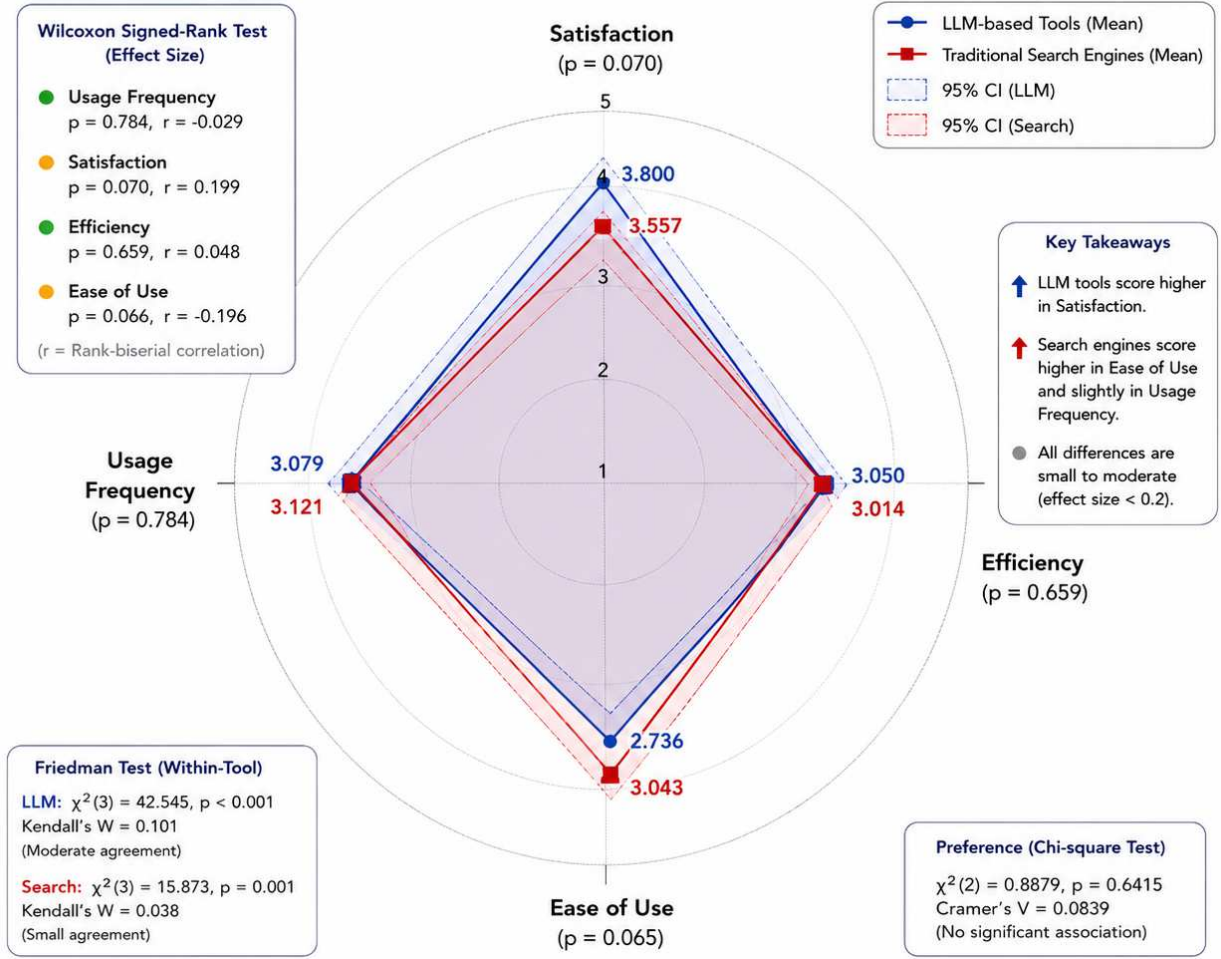


Figure 4: Overall comparison of LLM-based tools and traditional search engines across the four evaluated dimensions. Solid lines represent mean Likert scores, shaded regions indicate 95% confidence intervals, and annotations summarize the Wilcoxon signed-rank, Friedman, and chi-square test results.

For LLM-based tools, the Friedman test revealed a statistically significant difference among usage frequency, satisfaction, efficiency, and ease-of-use ratings, $\chi^2(3) = 42.545$, $p < .001$, with a small effect size ($W = .101$). A statistically significant difference was also observed among the four measures for traditional search engines, $\chi^2(3) = 15.873$, $p = .001$, although the effect size was very small ($W = .038$).

These results indicate that participants' ratings varied across the different dimensions within each tool category. However, the Friedman test alone does not identify which specific pairs of dimensions differ from one another. Therefore, these findings should be interpreted as evidence of overall within-tool variation rather than evidence that one tool category was superior to the other.

4.5 Association Between Gender and Overall Tool Preference

To examine whether gender was associated with overall tool preference, a chi-square test of independence was conducted. The analysis did not identify a statistically significant association between gender and overall tool preference, $\chi^2(2) = 0.888$, $p = .642$. The corresponding Cramer's V was 0.084, indicating a negligible association.

Therefore, within this sample, the distribution of overall preference among LLM-based tools, traditional search engines, and hybrid use did not differ significantly by gender.

Table 3: Paired comparisons between LLM-based tools and traditional search engines.

Measure	LLM Mean	Search Mean	W	<i>p</i>
Usage Frequency	3.079	3.121	3352.5	.784
Satisfaction	3.800	3.557	2230.0	.070
Efficiency	3.050	3.014	2960.5	.659
Ease of Use	2.736	3.043	2683.0	.066

Table 4: Friedman test results for repeated measures within each tool category.

Tool Category	χ^2	df	<i>p</i>	Kendall’s <i>W</i>
LLM-based tools	42.545	3	< .001	.101
Traditional search engines	15.873	3	.001	.038

4.6 Qualitative Analysis

The qualitative analysis draws on participants’ open-ended survey responses and in-person interview transcripts to examine how university students perceive, experience, and strategically select between LLM-based tools and traditional search engines when preparing for English language tests for higher studies. A thematic analysis approach was employed to identify recurring patterns across the qualitative data. Two researchers independently conducted initial line-by-line coding of all textual responses. These codes were subsequently discussed, compared, and iteratively refined into higher-level themes until consensus was achieved, ensuring analytical rigor and reliability.

Four dominant themes emerged from the analysis: (1) task suitability and tool selection, (2) perceptions of reliability and accuracy, (3) workflow efficiency and cognitive load, and (4) usability and interaction experience.

Task Suitability and Tool Selection. Participants consistently differentiated between LLMs and search engines based on the nature of the English language test task. LLM-based tools were frequently described as particularly useful for tasks such as summarization, paraphrasing, grammar correction, and drafting short written responses. Participants noted that LLMs were especially helpful under time constraints, as they provided immediate and structured outputs that could be refined further. In contrast, traditional search engines were preferred for tasks requiring detailed rule verification, exposure to multiple examples, or consultation of authoritative sources. Several participants emphasized that search engines were more suitable when confirming grammar rules or reviewing multiple perspectives before finalizing an answer.

Perceptions of Reliability and Accuracy. Trust emerged as a central factor influencing tool choice. While participants appreciated the fluency and coherence of LLM-generated responses, many expressed reservations regarding their factual reliability, particularly for nuanced grammar rules or language-specific explanations. Several respondents reported instances in which LLM outputs appeared convincing but required additional verification. Consequently, traditional search engines were frequently used as a secondary validation mechanism. Participants generally viewed search engines as more dependable for fact-checking and accessing credible sources, although some noted difficulties in evaluating source quality or resolving contradictory information.

Workflow Efficiency and Cognitive Load. LLM-based tools were widely perceived as effective in reducing cognitive effort, particularly under time pressure. Participants highlighted that LLMs helped them avoid navigating multiple web pages by providing concise explanations or draft responses in a single interaction. However, students who combined both tools acknowledged that switching between LLMs and

search engines increased overall task duration. Despite this, they reported that the additional effort often led to improved understanding and higher-quality outputs. This hybrid workflow was especially common for complex English language test tasks involving structured writing, reading comprehension, or vocabulary usage.

Usability and Interaction Experience. Participants frequently described LLMs as interactive and supportive, likening them to a conversational tutor that guided them through problem-solving steps. This conversational interaction was perceived as particularly beneficial for clarifying misunderstandings and exploring alternative phrasing. In contrast, traditional search engines were characterized as familiar and stable but less interactive, requiring greater user initiative to locate and synthesize relevant information. Interface familiarity, preferred learning style, and comfort with open-ended search influenced tool preference, especially among students with lower confidence in English language proficiency.

Overall, participants did not view LLM-based tools and traditional search engines as direct substitutes but rather as complementary components within their English language test preparation workflow. LLMs were valued for their speed, language generation capabilities, and explanatory support, while search engines remained essential for verification, rule confirmation, and access to credible sources. Tool selection was highly context-dependent, shaped by task type, prior knowledge, and the user’s need for either efficiency or accuracy. These findings highlight the nuanced strategies students adopt when navigating digital information tools and underscore the importance of integrated systems that balance convenience with reliability.

4.6.1 Reliability Analysis

The internal consistency of the survey instrument was assessed using Cronbach’s alpha. The overall reliability of the scale was $\alpha = 0.81$, indicating good internal consistency. Subscale reliability values were also satisfactory, with usage frequency ($\alpha = 0.78$), satisfaction ($\alpha = 0.83$), efficiency ($\alpha = 0.74$), and ease of use ($\alpha = 0.86$). These results confirm the reliability of the survey measures used in this study.

5 Discussion

The combined findings from the survey and in-person interviews reveal a complex and context-sensitive relationship between LLM-based tools and traditional search engines in English language test problem-solving for higher studies. Rather than exhibiting a clear preference for a single tool, students demonstrated adaptive strategies that varied according to task requirements, time constraints, and confidence in the information needed.

LLM-based tools were widely appreciated for their ability to generate structured, coherent, and contextually appropriate responses. Participants frequently highlighted the usefulness of conversational interfaces for tasks such as summarization, paraphrasing, grammar correction, and short-form writing, where rapid feedback and reduced cognitive effort were particularly beneficial. These strengths were reflected in higher usability ratings and faster task completion times for LLM-only users. However, despite these advantages, participants remained cautious about relying solely on LLMs due to concerns regarding occasional inaccuracies, generalized explanations, or the absence of explicit source attribution. As a result, many students reported verifying LLM-generated outputs through external sources.

Traditional search engines, most notably Google, continued to play a critical role in tasks that demanded depth, credibility, and source verification. Participants emphasized the importance of accessing authoritative websites, academic resources, and multiple viewpoints, particularly for grammar rules, reading comprehension, and fact-based validation. At the same time, students reported challenges associated with traditional search engines, including information overload, fragmented content, and the cognitive effort required to compare and synthesize information across multiple links. These limitations often led to longer task completion times and increased mental effort.

A central insight emerging from both quantitative and qualitative data was students’ strong inclination toward a hybrid approach that integrates the strengths of both LLMs and traditional search engines. As illustrated in Figure 5, the majority of participants expressed a preference for using both tools in combination rather than relying exclusively on either one. This preference reflects a desire for systems that support efficient content generation while maintaining access to credible, verifiable information sources.

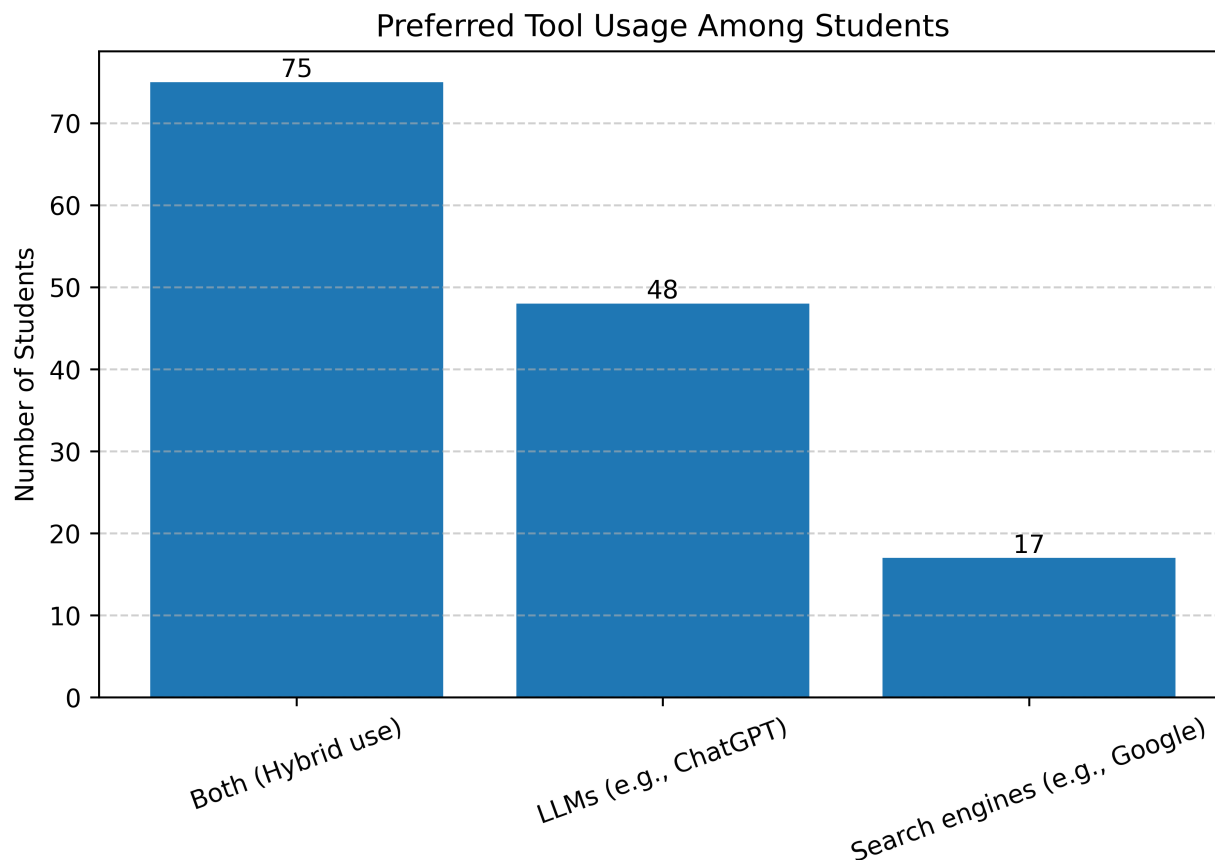


Figure 5: Distribution of participants’ overall tool preference for English language test problem solving. The majority of students preferred a hybrid approach combining LLM-based tools and traditional search engines, followed by exclusive use of LLMs, while comparatively few relied solely on traditional search engines.

Evidence from the interview study further reinforces the value of this integrated strategy. Participants who adopted a tool-balancing approach using LLMs for rapid explanation and drafting and search engines for verification and refinement achieved the highest task accuracy (90%). Although this approach required more time than using a single tool, participants consistently viewed the additional effort as a worthwhile trade-off. They reported greater confidence in their final responses, improved understanding of the material, and higher overall output quality. In contrast, participants who relied exclusively on either LLMs or search engines demonstrated lower accuracy, highlighting the limitations of single-tool reliance.

Taken together, these findings suggest that students do not perceive LLMs and traditional search engines as competing technologies but rather as complementary components of an effective academic workflow. LLMs excel in synthesis, explanation, and language generation, while search engines provide depth, reliability, and access to authoritative sources. The strategic integration of these capabilities represents a promising direction for the design of future academic support systems.

More broadly, the results underscore the importance of developing intelligent, context-aware hybrid tools that reduce cognitive load without compromising academic rigor. By minimizing unnecessary tool-switching and embedding verification mechanisms within generative interfaces, such systems have the potential to enhance learning efficiency, accuracy, and trust in AI-assisted academic problem solving.

5.1 Practical Implications

The findings suggest that LLM-based tools are particularly effective for writing, paraphrasing, and summarization tasks, while traditional search engines remain valuable for source verification and accessing author-

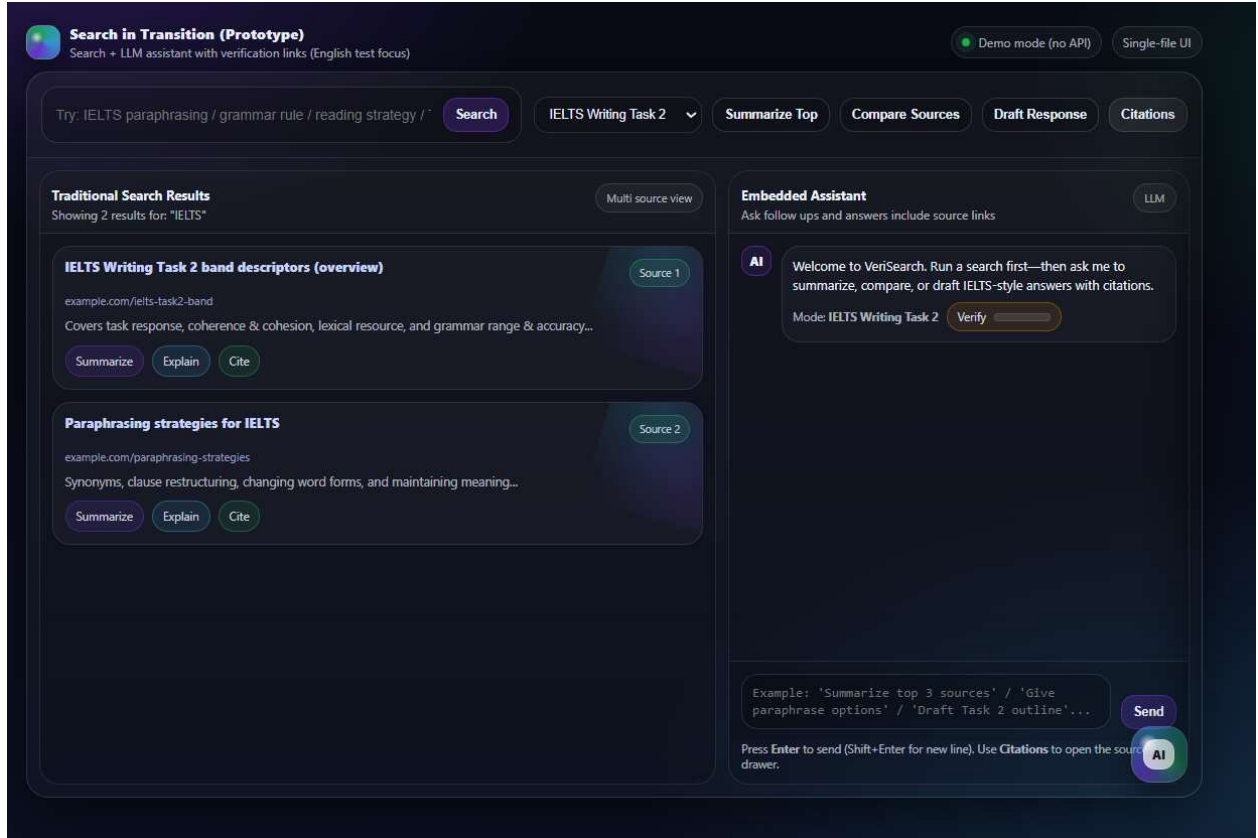


Figure 6: Conceptual prototype of the proposed hybrid search-LLM interface for English language test preparation. The interface integrates traditional search results with an embedded LLM assistant that provides explanations, summaries, and draft responses, while explicitly linking generated outputs to underlying sources to support verification and reduce cognitive load.

itative references. A hybrid strategy that combines both tools appears to provide the best balance between efficiency and accuracy for English language test preparation.

5.2 Proposed Prototype

Drawing on user insights from both the survey responses and the in-person interviews, we propose a conceptual, HCI-oriented hybrid prototype that embeds LLM-based assistance directly within a traditional search engine interface. The primary objective of this design is to reduce cognitive effort and workflow fragmentation caused by frequent tool switching by offering a unified interaction environment that combines the conversational utility of LLMs with the source-rich infrastructure of search engines. The design specifically targets English language test problem-solving for higher studies, reflecting the dominant use cases identified in this study.

The prototype is envisioned as an unobtrusive embedded chatbot positioned within the search interface, allowing users to engage in context-aware dialogue without disrupting their familiar browsing behavior. Unlike standalone LLM interfaces, the assistant does not replace standard search results. Instead, it complements them by offering concise summaries, follow-up clarifications, and cross-source synthesis derived from retrieved content. For example, after a query related to English language test preparation, the assistant can summarize the top search results, compare key points across multiple sources, and simplify explanations for grammar rules, reading passages, or writing tasks. Users may then ask follow-up questions to refine or extend the information without repeatedly navigating and interpreting multiple webpages.

A key contribution of the design lies in its hybrid interaction model, which enables users to move seam-

lessly between raw search content and AI-supported interpretation. To address concerns regarding LLM trustworthiness raised by participants, each AI-generated response is explicitly accompanied by links to the underlying sources. This design choice promotes transparency, supports verification, and mitigates the risk of unverified or hallucinated outputs in academic contexts.

Although conceptual in its current form, the prototype is grounded in the empirical findings of this study and reflects users' demonstrated preference for combining LLM-based tools and traditional search engines. The prototype is intended to illustrate interaction design concepts rather than serve as a fully implemented system. By embedding an assistive layer into the search experience, the design aims to reduce cognitive load, increase task efficiency, and encourage evidence-based academic practices by bridging the gap between content generation and source verification. Figure 6 shows the conceptual prototype, and a demonstration version of the conceptual prototype interface is publicly [available](#) to support transparency and reproducibility.

6 Limitations and Future Work

This study has a number of limitations. To begin with, the quantitative survey garnered a fairly large sample (140 University students) while the qualitative component of the research had only 20 interview participants. This is an acceptable sample size as it is the qualitative findings that are not transferable. With future research of larger and more varied samples of interviews, there would be more evidence for the students' interaction strategies and perceptions.

Secondly, most of the participants belonged to the same University of Bangladesh, and a good portion of them were from the technical streams such as Computer Science and Engineering (CSE), etc. Therefore, the use, behaviour, and attitudes reported and expressed may not necessarily apply to the students of other institutions, disciplines of study, or cultural and educational backgrounds. The findings need to be more broadly generalizable through multi-institutional/ cross cultural investigations.

Third, measurements used for quantitative analyses were largely self-reported usage frequency, satisfaction, perceived efficiency, and ease of use. The instrument used to collect the data in this survey had good internal consistency, as the Cronbach's alpha value was 0.81; however, self-reported data can be vulnerable to recall bias, response bias, and social desirability bias. The study findings suggest that further research should be conducted that combines the survey findings with other objective behavioral data, e.g., through interaction logs, clickstream data, eye tracking data, or screen recording data, to fully understand students' information-seeking behaviors in the future.

Fourth, pre-made rubrics for each task were used in evaluating the interview results, and the tasks were similar to typical IELTS and TOEFL preparation tasks, which were created with teacher input and assistance in English. However, these controlled tasks can at best only reflect a part of an actual L2, long-term, English-learning experience. Future research in this study would be beneficial to execute longitudinal research in authentic education environments, to gain more insight into the student's tool-selection strategies over time.

The qualitative data were then independently coded by two coders, with discussion and disagreements until goals were reached, although a quantitative inter-coder reliability measure (e.g., Cohen's κ) was not obtained. For future qualitative investigation, inter-coder agreement statistics should be included in and reported to meet methodological rigor and methodological reproducibility.

Finally, the computational model proposed in this paper – integration of search and LLM functions – is an idea that has not been implemented and tested for empirical results as of yet. Future research might be done in order to develop and validate the functional prototype by conducting controlled usability studies (or experiments in real classrooms) which could focus on the aforementioned categories of cognitive load, learning performance, trust, user satisfaction, and long-term retention of knowledge. Moreover, the role of the learner factors of experience in LLM, digital literacy, and English language proficiency and information retrieval ability in future interaction strategies and learning outcomes in AI-assisted educational environments should be examined.

7 Data Availability

All the data that is referenced or discussed in the current study, including the anonymized survey data, survey instrument, interview protocol, task evaluation rubric, variable descriptions, and resources for statistical analyses, are publicly available in the GitHub repository: [Supplementary Materials](#). All the data shared are anonymous to protect the confidentiality of participants. Supporting materials for reproducing analyses shown in this study are also included in the repository.

8 Conclusion

This paper was a mixed-method study on the employment of large language model (LLM) tools compared to traditional (non-LLM) search engines for English language exam preparation among university students pursuing higher studies. The study features both quantitative and qualitative evidence gathered from a survey of 140 students and task-based interviews with 20 students, offering a holistic picture of students' strategies of utilizing conversation AI and traditional information retrieval in the process of tackling English language test problems.

To assess significant differences between LLM tools and traditional search engines descriptively, quantitative results revealed differences between the two for the frequency of use, level of user satisfaction, perceived efficiency, and ease of use. However, paired statistical comparisons revealed that there were no significant differences between the two pairs of tools in terms of certain learning purposes and therefore both strategies were regarded as beneficial in the students' eyes. The results of the qualitative analysis additionally revealed that the choice of tools is very context-sensitive. Overall, most participants tended to choose conversational tools to find answers, including summarisation, paraphrasing, explanation of grammar rules, and written answers, because they were easy to use and fast to get answers, while traditional search engines were familiar, easy to use, and trusted to find authoritative information or to verify sources.

The results of the interview-based evaluation also showed that the "balanced hybrid" group that used both the LLM-based tools and the traditional search engines had the highest mean task accuracy (90%), followed by the "LLM-only" group (88%) and the "traditional search engine-only" group (80%). Overall, the mean was higher for the hybrid condition for the time it took to complete the tasks, but in both, respondents were more confident of the information provided, that is, the correctness and validity of the information gathered to make an answer.

This study, with its empirical results, proposes a theoretical, human-computer interaction (HCI) design method to create a hybrid interface, mixing conversational LLM support and traditional search interfaces. Explanations, summarisation, and giving credit to information provided by a source are also included in the proposed design, supported by the unifier of using AI, to eliminate the fragmentation of the workflow, verify information, and give trust to users in using AI-supported learning.

The findings demonstrate that LLM-based tools do not compete with traditional search engines as tools for L2 English test preparation but can be used as complementary tools. Conversational AI can enhance information retrieval's effectiveness, transparency, and trustworthiness, making their integration more effective for creating better learning experiences. The hybrid interaction model introduced in this work provides implications for implementing future learning systems that support the integration of generative AI with conventional search methods, and it lays the groundwork for future research into human-centric learning environments with generative AI and conventional search capabilities.

References

- [1] U. Alberth, "The use of ChatGPT in academic writing: A blessing or a curse in disguise?", *TEFLIN Journal*, vol. 34, pp. 337–352, 2023.
- [2] S. Bal and R. Nath, "A comparative study of traditional search engines with the metasearch engines," *Ultra Scientist*, vol. 21, no. 2M, pp. 597–610, 2009.

- [3] A. Bansal, “Optimizing RAG with hybrid search and contextual chunking,” *Journal of Emerging Applications in Science and Technology*, vol. 5, 2023.
- [4] K. M. Caramancion, “Large language models vs. search engines: Evaluating user preferences across varied information retrieval scenarios,” *arXiv preprint arXiv:2401.05761*, 2024.
- [5] R. R. Divekar, S. Guerra, L. Gonzalez, and N. Boos, “Choosing between an LLM versus search for learning: A higher-ed student perspective,” *arXiv preprint arXiv:2409.13051*, 2024.
- [6] P. Kapoor, S. Mahida, S. John, *et al.*, “AI-driven adaptive systems for personalized library research assistance,” *Library of Progress: Library Science, Information Technology & Computer*, vol. 44, no. 3, 2024.
- [7] H. Kumar, M. Reza, J. Mitchell, I. Musabirov, L. Zhang, and M. Liut, “Understanding help-seeking behavior of students using LLMs vs. web search for writing SQL queries,” *arXiv preprint arXiv:2408.08401*, 2024.
- [8] Perplexity AI, “Perplexity AI: Conversational search engine,” 2024, accessed Nov. 11, 2024.
- [9] F. A. Pirzado, A. Ahmed, R. A. Mendoza-Urdiales, and H. Terashima-Marin, “Navigating the pitfalls: Analyzing the behavior of LLMs as a coding assistant for computer science students,” *IEEE Access*, 2024.
- [10] T. Sakirin and R. Ben Said, “User preferences for ChatGPT-powered conversational interfaces versus traditional methods,” *Mesopotamian Journal of Computer Science*, pp. 22–28, 2023.
- [11] S. E. Spatharioti, D. M. Rothschild, D. G. Goldstein, and J. M. Hofman, “Comparing traditional and LLM-based search for consumer choice: A randomized experiment,” *arXiv preprint arXiv:2307.03744*, 2023.
- [12] A. Wazzan, S. MacNeil, and R. Souvenir, “Comparing traditional and LLM-based search for image geolocation,” in *Proc. 2024 Conf. on Human Information Interaction and Retrieval*, pp. 291–302, 2024.
- [13] R. Xu, Y. Feng, and H. Chen, “ChatGPT vs. Google: A comparative study of search performance and user experience,” *arXiv preprint arXiv:2307.01135*, 2023.
- [14] E. Kasneci *et al.*, “ChatGPT for good? On opportunities and challenges of large language models for education,” *Learning and Individual Differences*, vol. 103, p. 102274, 2023.
- [15] Rahman, T., Hossen, M. S., Mondol, M. P., & Mukta, J. N. (2026). *Search in Transition: A study of university students’ perspectives on using LLMs and traditional search engines in English test problem solving for higher study*. arXiv preprint arXiv:2602.13629. Available at: <https://arxiv.org/abs/2602.13629>

Acknowledgments

The authors gratefully acknowledge all authors who voluntarily participated in the online surveys and interviews. Their valuable time, thoughtful responses, and meaningful contributions were essential to the successful completion of this study.

Ethics, Consent to Participate, and Consent to Publish

Ethics approval: This research was conducted under the ethical approval granted by the Institutional Review Board (IRB) of BRAC University, Dhaka, Bangladesh (Approval No. BRACUIRB.220240006; approved on 06 May 2024). The study formed part of the research activities covered by this IRB approval. All research procedures involving human participants complied with the ethical guidelines established by

the Institutional Review Board and were carried out in accordance with the principles of the Declaration of Helsinki and its subsequent amendments.

Consent to participate: Prior to participation, informed consent was obtained from all participants. Individuals who completed the online survey provided electronic written consent, whereas participants in the interviews gave verbal informed consent before the interviews commenced. Participation was entirely voluntary, and all participants were informed of their right to withdraw from the study at any stage without any adverse consequences.

Consent to publish: Not applicable. The manuscript does not include any personally identifiable information or images of participants that would require consent for publication.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

AI Use Disclosure

The authors used ChatGPT (OpenAI) exclusively for language editing and grammatical refinement during manuscript preparation. No AI tools were used for data analysis, interpretation of results, or content generation.

A Survey Instrument

Survey Instrument Overview

All survey items were administered in English and were reviewed for clarity prior to distribution. The questionnaire examined university students' use and perceptions of Large Language Models (LLMs) and traditional search engines in the context of English language test preparation.

A.1 Study Information

This survey was conducted as part of an academic research study examining how university students use Large Language Models (LLMs) and traditional search engines to solve English language test problems for higher studies, including IELTS and TOEFL preparation.

The study focuses on user interaction patterns, usability, efficiency, and perceived accuracy of NLP-based systems in English test preparation and problem-solving contexts.

Participation was voluntary, and responses were collected for academic research purposes.

A.2 Demographic Information

Participant Background

1. What is your birth year range?

- 18–20 years
- 21–25 years
- 26–30 years

2. What is your gender?

- Female
- Male
- Prefer not to say
- Other

3. Please select your current CGPA range (on a 4.00 scale).

- 3.81–4.00
- 3.51–3.80
- 3.01–3.50
- 2.50–3.00

A.3 Participation and English Test Preparation

Participation and Test Preparation

1. I voluntarily agree to participate in this study.
 - Yes, I agree
2. Which English language test are you preparing for?
 - IELTS
 - TOEFL
 - Duolingo
 - Planning to take one
 - Other

A.4 Tool Usage Frequency

Participants were asked to report how frequently they used LLM-based tools and traditional search engines for English test preparation.

A.4.1 LLM-Based Tools

Question: LLM Usage Frequency

How often do you use LLMs (e.g., ChatGPT) for English test preparation?

- Never
- Rarely
- Occasionally
- Sometimes
- Often
- Always

A.4.2 Traditional Search Engines

Question: Search Engine Usage Frequency

How often do you use Google or other search engines for English test preparation?

- Never
- Rarely
- Occasionally
- Sometimes
- Often
- Always

A.5 Perceived Accuracy and Satisfaction

A.5.1 LLM-Based Tools

Question: LLM Accuracy and Satisfaction

How satisfied are you with the accuracy of LLMs in solving English test problems?

- Very Dissatisfied
- Dissatisfied
- Neutral
- Satisfied
- Very Satisfied

A.5.2 Traditional Search Engines

Question: Search Engine Accuracy and Satisfaction

How satisfied are you with the accuracy of search engines in solving English test problems?

- Very Dissatisfied
- Dissatisfied
- Neutral
- Satisfied
- Very Satisfied

A.6 Perceived Efficiency

A.6.1 LLM-Based Tools

Question: LLM Efficiency

How efficient are LLMs in helping you complete English test tasks?

- Not Efficient at All
- Slightly Efficient
- Moderately Efficient
- Very Efficient
- Extremely Efficient

A.6.2 Traditional Search Engines

Question: Search Engine Efficiency

How efficient are search engines in helping you complete English test tasks?

- Not Efficient at All
- Slightly Efficient
- Moderately Efficient
- Very Efficient
- Extremely Efficient

A.7 Ease of Use

A.7.1 LLM-Based Tools

Question: LLM Ease of Use

How easy is it to use LLMs for English test problem solving?

- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

A.7.2 Traditional Search Engines

Question: Search Engine Ease of Use

How easy is it to use search engines for English test problem solving?

- Very Difficult
- Difficult
- Neutral
- Easy
- Very Easy

A.8 Overall Tool Preference

Question: Overall Tool Preference

Which tool do you prefer overall for English test preparation?

- LLMs (e.g., ChatGPT)
- Search engines (e.g., Google)
- Both (Hybrid use)

A.9 Open-Ended Questions

Qualitative Questions

1. In which English test tasks (e.g., reading, writing, grammar, vocabulary) do you prefer using LLMs over search engines, and why?
2. What limitations or challenges have you faced while using LLMs or search engines for English test preparation?

B Interview Tasks

Assigned English Language Test-Based Tasks

Participants were asked to complete the following English language test-related tasks using GPT-based tools, traditional search engines, or both:

1. **Summarizing an english languages test based passage.** Participants were asked to read english languages test based passage and produce a concise summary that preserved the main arguments and key supporting points while maintaining perfect tone.
2. **Correcting grammatical errors in a short text.** Participants reviewed languages testing passage containing deliberate grammatical and syntactic errors and were instructed to identify and correct these errors in accordance with standard English conventions.
3. **Selecting appropriate vocabulary in context.** Participants selected contextually appropriate vocabulary items to complete sentences, focusing on lexical accuracy, collocation, and semantic appropriateness commonly assessed in English language proficiency tests.
4. **Paraphrasing sentences without altering meaning.** Participants paraphrased given sentences while preserving the original meaning, emphasizing lexical variation, syntactic restructuring, and avoidance of plagiarism, as required in academic writing and standardized tests.
5. **Drafting an opinion-based response similar to IELTS Writing Task 2.** Participants composed a argumentative response to a prompt resembling IELTS Writing Task 2, focusing on coherence, clarity of argument, lexical range, and grammatical accuracy.
6. **Drafting a response similar to IELTS Writing Task 1.** Participants wrote a brief descriptive response based on visual or textual information, mirroring IELTS Writing Task 1 requirements, with emphasis on accurate data interpretation, organization, and appropriate use of academic language.
7. **Interpreting information from an academic reading passage.** Participants answered comprehension and interpretation questions based on an academic reading passage, requiring identification of main ideas, inference, and understanding of supporting details.

C Thematic Codebook

Thematic analysis of the open-ended survey responses and interview data identified four overarching themes related to students' experiences with LLM-based tools and traditional search engines in English language test preparation. These themes capture differences in perceived effectiveness, usability, trust and verification, and context-dependent tool selection. Each overarching theme was further organized into child codes representing recurring patterns in participants' responses. Table 5 presents the thematic structure, code definitions, and illustrative participant statements.

Parent Theme	Child Code	Definition	Illustrative Participant Statement
Effectiveness	Task Fit	Perceived suitability of LLM-based tools or traditional search engines for specific English language test preparation tasks, including writing, grammar, reading, vocabulary, paraphrasing, and summarization.	“LLMs are useful for writing and explanations, while search engines are helpful when I need specific examples or additional sources.”
Usability and Cognitive Load	Ease of Use	The perceived simplicity, convenience, and intuitiveness of using a tool during English language test preparation.	“ChatGPT is easier because I can ask a question directly and receive an answer without searching through many different websites.”
	Information Overload	Difficulty caused by excessive, fragmented, or irrelevant information, particularly when users must examine multiple search results to identify useful material.	“Search engines provide many results, and sometimes it takes time to find the information that is actually useful.”
Trust and Credibility	Source Verification	The perceived need to verify information, explanations, grammar rules, or writing guidance using reliable sources, particularly when accuracy and credibility are important.	“I prefer to check information from search results when I need to confirm whether an explanation or grammar rule is correct.”
	Perceived Accuracy	Users’ perceptions of the correctness, reliability, and consistency of information or answers provided by LLM-based tools and traditional search engines.	“LLMs are convenient, but I sometimes verify the answer because the information may not always be completely accurate.”
Contextual Selection	Tool Task Type Influence	The selection of an LLM-based tool, traditional search engine, or both depending on the type, complexity, and requirements of the English language test task.	“I use an LLM for writing and grammar questions, but I use search engines when I need examples, sources, or additional explanations.”
	Hybrid Use	The complementary use of LLM-based tools and traditional search engines to combine direct assistance with information search and verification.	“I usually use both tools because the LLM gives a quick explanation, and search engines help me verify the information.”

Table 5: Thematic codebook developed from the qualitative responses collected in the study. The codebook summarizes the overarching themes, child codes, operational definitions, and illustrative participant statements concerning the use of LLM-based tools and traditional search engines for English language test preparation.

Supplementary Files

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

- [analysis.py](#)
- [InterviewData.xlsx](#)
- [SearchinTransitionfinalversione.xlsx](#)
- [SurveyQusetion.htm](#)
- [STATISTICSanalysis.ipynb](#)