LingoQ: Bridging the Gap between EFL Learning and Work through AI-Generated Work-Related Quizzes

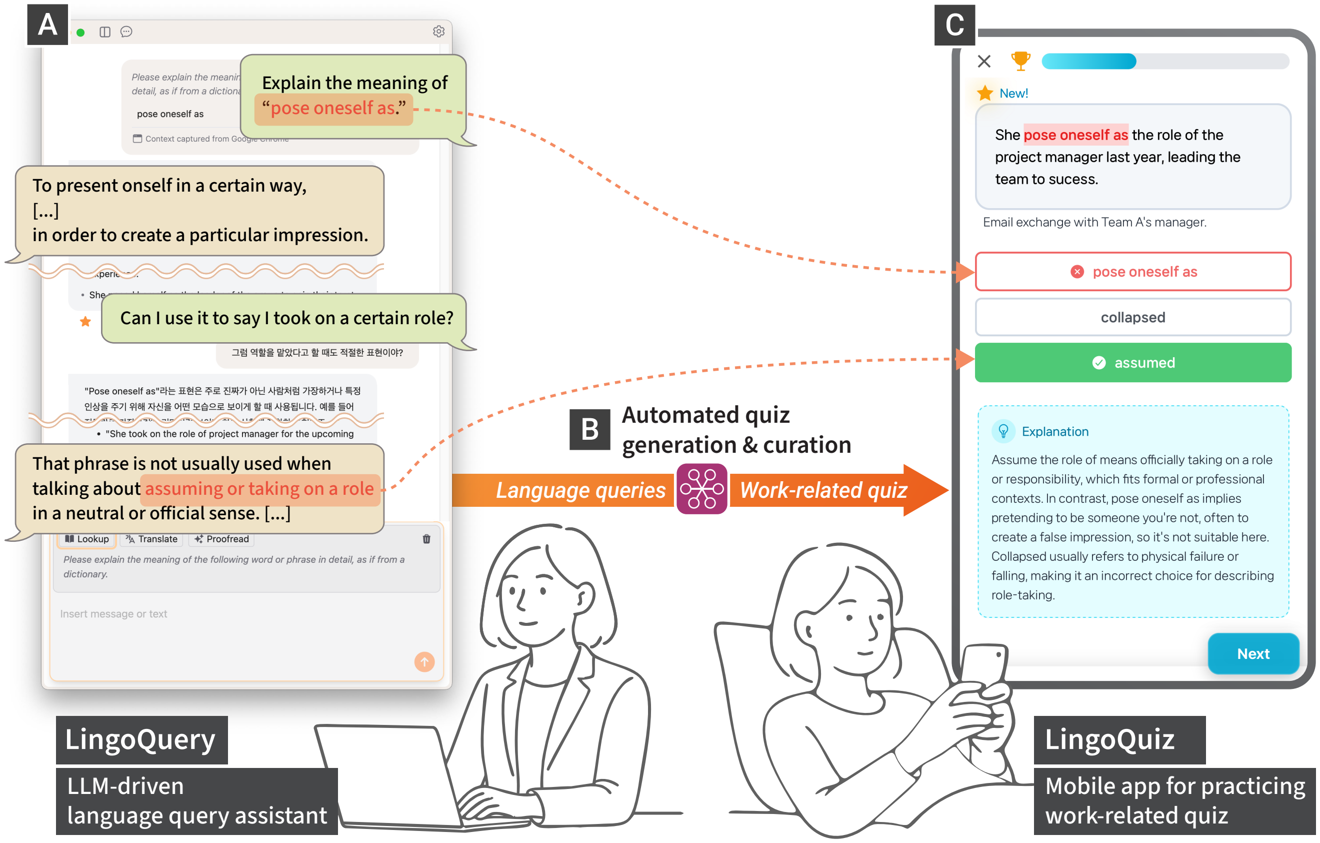
Yeonsun Yang[*]
DGIST
Republic of Korea
diddustjs98@dgist.ac.kr

Sang Won Lee[†]
Virginia Tech
Blacksburg, VA, USA
sangwonlee@vt.edu

Jean Y. Song
Yonsei University
Republic of Korea
jeansong@yonsei.ac.kr

Sangdoo Yun
NAVER AI Lab
Republic of Korea
oodgnas@gmail.com

Young-Ho Kim
NAVER AI Lab
Republic of Korea
yghokim@younghokim.net


A
Explain the meaning of "pose oneself as."
To present onself in a certain way, [...] in order to create a particular impression.
Can I use it to say I took on a certain role?
That phrase is not usually used when talking about assuming or taking on a role in a neutral or official sense. [...]
B
Automated quiz generation & curation
Language queries
Work-related quiz
C
LingoQuery
LLM-driven language query assistant
LingoQuiz
Mobile app for practicing work-related quiz

Please explain the meaning of the following word or phrase in detail, as if from a dictionary.

pose oneself as

Context captured from Google Chrome

그럼 역할을 맡았다고 할 때도 적절한 표현이야?

"Pose oneself as"라는 표현은 주로 진짜가 아닌 사람처럼 가장하거나 특정 인상을 주기 위해 자신을 어떤 모습으로 보이게 할 때 사용됩니다. 예를 들어

- "She took on the role of project manager for the upcoming

Lookup | Translate | Proofread

Insert message or text

New!

She pose oneself as the role of the project manager last year, leading the team to sucess.

Email exchange with Team A's manager.

- pose oneself as
- collapsed
- assumed

Explanation

Assume the role of means officially taking on a role or responsibility, which fits formal or professional contexts. In contrast, pose oneself as implies pretending to be someone you're not, often to create a false impression, so it's not suitable here. Collapsed usually refers to physical failure or falling, making it an incorrect choice for describing role-taking.

Next

Figure 1: LingoQ consists of three components. In LingoQuery for desktop A, information workers can interact with an LLM-based chatbot for English-related language queries. The automated quiz generation pipeline B produces and curates multiple choice English questions using the query interactions of LingoQuery as materials. In LingoQuiz C on a smartphone, workers can later review their language queries by completing the generated quizzes. (Please refer to our supplementary video, available at `https://naver-ai.github.io/lingo-q`, which demonstrates the interactions.)

Abstract

Non-native English speakers performing English-related tasks at work struggle to sustain EFL learning, despite their motivation. Often, study materials are disconnected from their work context. Our formative study revealed that reviewing work-related English becomes burdensome with current systems, especially after work. Although workers rely on LLM-based assistants to address their immediate needs, these interactions may not directly contribute to their English skills. We present LingoQ, an AI-mediated system that allows workers to practice English using quizzes generated from their LLM queries during work. LingoQ leverages these on-the-fly

[*] Yeonsun Yang conducted this work as a research intern at NAVER AI Lab.
[†] Sang Won Lee conducted this work as a visiting scholar at NAVER AI Lab.

CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791342

queries using AI to generate personalized quizzes that workers can review and practice on their smartphones. We conducted a three-week deployment study with 28 EFL workers to evaluate LingoQ. Participants valued the quality-assured, work-situated quizzes and constantly engaging with the app during the study. This active engagement improved self-efficacy and led to learning gains for beginners and, potentially, for intermediate learners. Drawing on these results, we discuss design implications for leveraging workers' growing reliance on LLMs to foster proficiency and engagement while respecting work boundaries and ethics.

CCS Concepts

• **Human-centered computing → Natural language interfaces**; **Empirical studies in HCI**.

Keywords

English as a foreign language, information workers, large language model, question generation, context awareness

ACM Reference Format:
Yeonsun Yang, Sang Won Lee, Jean Y. Song, Sangdoo Yun, and Young-Ho Kim. 2026. LingoQ: Bridging the Gap between EFL Learning and Work through AI-Generated Work-Related Quizzes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3772318.3791342

1 Introduction

In the global economy, where online resources are predominantly in English, information workers[1] who are not native English speakers often need English proficiency for their jobs. This requirement includes understanding English texts used in papers, articles, and reports, as well as the ability to communicate via email. To enhance their English proficiency, information workers frequently engage in self-directed language learning using mobile applications like Duolingo [33], Babbel [7], Memrise [77], and RosettaStone [92] that offers access to learning anytime, anywhere through mobile phones [105]. However, the mobile language learning apps often create study materials from generic situations like traveling or business meetings, limiting the depth and practical relevance of vocabulary and conversational skills. Such questions, based on ordinary scenarios, make it difficult to acquire the English skills needed for work-related tasks directly. For example, a programmer requiring English for API documentation gains little from fill-in-the-blank questions about family composition or reading comprehension exercises on threats to coral reefs in the Pacific Ocean.

To address the challenge of English studies being disconnected from workers' everyday tasks, we draw on task-based language teaching [81] and situated learning [68]. When study materials are embedded in workers' job contexts, they can improve task performance while sustaining engagement in language learning [3]. Prior research shows that grounding instruction in authentic tasks not only supports second language acquisition [75, 108], but also enhances motivation [55] and strengthens memory retention [21]. Building on these insights, we investigate whether generating study materials of mobile language learning apps directly from work tasks can further enhance language learning and foster long-term engagement. Prior work has explored context-aware content generation using daily situations, objects, or personal interests, and demonstrated technological feasibility and learning benefits for general learners [32, 37, 52, 111]. However, little has been explored how to support workers in EFL learning, even though a large amount of personal data from information-work contexts remains underutilized despite its potential to generate personalized learning materials.

To understand EFL workers' current strategies and challenges across both their work and learning environments, we conducted a formative study consisting of an online survey and follow-up interviews with 49 non-native information workers in South Korea. Participants expressed a strong willingness to improve their English proficiency for work, as lexical disruptions and low confidence in English frequently hindered their tasks. As a result, they turned to easily accessible English learning mobile apps, but the disconnection of the study materials from their work contexts reduced perceived helpfulness and engagement. In addition, participants heavily relied on LLM-based assistants, frequently asking to look up, translate, and proofread. To improve work-related English proficiency, a small subset of participants engaged in self-directed review practices: They manually recorded unfamiliar vocabulary and revisiting unclear passages, often transferring these materials into flashcard or quiz apps. However, participants reported that maintaining such manual practice routine was difficult, describing them as burdensome and taxing especially after work.

Based on the lessons from the formative study, we aimed to reduce learner burden and support engagement in work-related EFL learning by streamlining review practices of work-related English use, thereby shifting away from traditional EFL instructions—which often rely on decontextualized, curriculum-driven materials—toward a more contextualized, usage-driven pedagogical approach. As a first step toward this goal, we leveraged workers' interactions with LLM-based assistants for language queries—questions that EFL workers ask for their text-based English tasks and rarely revisit afterwards—to provide an automated routine for practice. We present LingoQ, an ensemble of intelligent systems designed to generate English quizzes from workers' LLM queries. LingoQ consists of three components: (1) LingoQuery, an LLM-based desktop chatbot that answers workers' English-related queries (A in Figure 1); (2) LingoQuiz, a mobile app that allows workers to complete short, context-relevant quizzes at their convenience (C in Figure 1); and (3) the backend pipeline that processes queries from LingoQuery to generate and validate quizzes (B in Figure 1). Leveraging conversation logs with LingoQuery as a source material, the backend automatically produces a set of personalized multiple-choice questions for each user using an LLM. The questions then undergo AI-driven quality checking and refinement to ensure that each has a single correct answer and an appropriate level of difficulty. On LingoQuiz, users can solve quizzes consisting of 10 questions that mix newly generated and previously solved ones.

We conducted a three-week field deployment with 28 information workers in South Korea. Through the deployment study, with a focus on feasibility, we explore the following research questions: **RQ1**–How do EFL workers engage in and sustain their English

[1] In this work, we use the term *information worker* to refer to workers whose primary job role involves gathering, synthesizing, and producing new information [65]. In this paper, we will use the term 'workers' to specifically refer to information workers.

learning when using study materials generated from their work context? **RQ2**–How does studying English with work-related content influence workers' learning outcomes and self-efficacy? and **RQ3**–How do EFL workers perceive the value of studying English with questions generated from work-related content?

Our results showed that participants consistently engaged with LingoQ throughout the study period and perceived review practice with LingoQ as more sustainable than their previous experiences with other study methods. Furthermore, our pipeline generated quality-assured questions that showed strong alignment with expert evaluations. We observed that participants' self-efficacy in English skills increased significantly after actively using LingoQ (9.5% gain on average, $p < 0.001$), and participants in the beginner-level English proficiency showed notable learning gain on a TOEIC-based English proficiency test. In the post-study survey, participants reported that quizzes generated by LingoQ were more relevant to their work and were helpful in their tasks compared to their prior experience of studying English. It also encouraged them to consider the educational aspect of their work-related queries.

The key contributions of this work are as follows:

(1) Design and implementation of LingoQ, an LLM-based system that supports lightweight review practice of daily English use by automatically generating quizzes from work-related English queries in an LLM-based assistant. LingoQ's design was informed by a formative study ($N = 49$) with non-native information workers. The source code of LingoQ is publicly available at `https://naver-ai.github.io/lingo-q`.

(2) Empirical findings from a three-week deployment study ($N = 28$) showing that LingoQ helps EFL workers sustain engagement and achieve measurable learning benefits through task-specific qualified quizzes with low burden, suggesting the potential of sourcing from on-the-fly LLM queries as an alternative to traditional decontextualized EFL learning methods.

(3) Design considerations for AI-mediated EFL learning material generation that fosters engagement and proficiency by leveraging knowledge workers' growing reliance on LLMs.

2 Related Work

In this work, we cover the related work in the areas of (1) EFL learning for information workers, (2) creating context-aware learning materials, and (3) retrieval practice for second language acquisition.

2.1 EFL Learning for Information Workers

English plays a critical role in the workplace as the common language of global industries and disciplines [63, 104]. The proliferation of digital work environments has posed unique challenges for non-native English-speaking workers, who navigate English-language texts as part of their computer-based tasks. For example, Amano *et al.* found that researchers spend 46.6% more time reading English papers and 50.6% more time writing them compared to native speakers, while facing higher rejection rates [1]. Similarly, non-native English programmers struggle with technical documentation, professional communication, and code comprehension [50]. Consequently, English-as-a-foreign-language (EFL) learning has become critical for information workers to develop English proficiency—the ability to achieve goals through the use of English in relation to the specific purposes [55].

In the field of HCI, previous research has explored approaches to directly support EFL workers' use of English in the workplace, such as assisting with email composition, providing on-demand evaluations of writing, simplifying complex texts, and paraphrasing with AI explanations, in order to reduce context-switching burdens and disruptions during work [15, 17, 24, 54, 59, 62, 66]. However, effective support for EFL workers in the long run requires fostering English skills that directly enhance their ability to accomplish work tasks. In response, our work focuses on cultivating English proficiency within professional contexts.

Conventional English education, with its emphasis on general language proficiency, often falls short of meeting the specific needs of professional contexts. To address this gap, education theories such as task-based language teaching [81] and situated learning [68] underscore the critical role of *authentic* materials—texts and resources originated from real-world application [43]. Authentic materials give learners up-to-date domain information, unlike predefined materials that may lag behind current developments [9]. Other studies found that working with content related to one's professional field enhances motivation and self-efficacy by building confidence in handling real-world tasks [10]. Context-specific, task-based materials help learners develop active knowledge directly applicable to their daily work [110]. Building on this emphasis, our work explores flexible ways to support EFL learning by providing authentic, work-related materials generated from workers' daily English tasks.

2.2 Creating Context-Aware Learning Materials

Context is fundamental to learning because knowledge is inseparable from the situations and activities in which it is acquired [19, 28]. Context-aware learning approaches [56], which situate learning in the learner's personal context by selecting, adapting, or generating content, are particularly well-suited for language learning. Because contextually relevant materials allow learners to experience language in authentic settings [70] and foster situational interests [53], they thereby enhance motivation and engagement [53], prevent inert knowledge [74], and ultimately promote meaningful and effective learning [32].

The field of HCI has explored context-aware personalized learning materials that adapt to factors such as learners' location [37, 52], surrounding elements [32, 57], social media content [18, 111], and other contextual information [83]. For example, MicroMandarin [37] suggested flashcards relevant to nearby venues, while Vocabura [52] generated L1-L2 word pairs from walking commute routes, both leveraging GPS coordinates to support vocabulary learning. Draxler *et al.* further explored an object-based approach that automatically generates exercises by detecting elements in learner-captured photos from their daily contexts [32]. In addition, Yamaoka *et al.* introduced a method that extracts keywords from Instagram posts to generate example sentences, helping learners acquire new words aligned with their interests and improving retention [111]. Beyond academic research, recent language learning services have begun adopting generative AI to provide situated and contextualized learning experiences [34, 45].

With the growth of digital environments, research has increasingly focused on usage-based learning to help learners acquire practical language skills by leveraging learner–computer interactions as sources of contextual content, such as eye gaze [29], clicked hyperlinks [25, 94], translations [76], and other digital traces [6]. Ding *et al.* identified unknown words through gaze trajectories while learners read foreign language texts, offering real-time translations and explanations for just-in-time vocabulary acquisition [29]. Lungu *et al.* proposed a comprehension approach that generated mobile exercises from learners' translated sentences during web reading, which could serve as potential learning cues [76]. This work demonstrated the feasibility of this ecosystem, highlighting engagement and learning benefits.

Our work extends this line of research by leveraging emerging conversational interactions between learners and LLM-based chatbots. We argue that queries to LLMs reflect learners' immediate language difficulties and learning intentions, serving as valuable cues for situated and usage-based learning. In this work, we aim to generate work-related learning exercises from EFL workers' queries collected in the course of their professional tasks.

2.3 Retrieval Practice for Second Language Acquisition

In second language acquisition, the *practice* of difficult linguistic features offers learners opportunities for meaningful language use, reinforces task performance, and fosters adaptive language proficiency [60, 102]. In particular, retrieval practice, which involves actively recalling knowledge from memory, typically through exercises such as quizzes or self-testing, is one of the most effective review strategies [93, 103]. It has been shown to be beneficial than simple 'restudy' in strengthening long-term memory and supporting the transfer of knowledge to new context [89, 93]. Accordingly, prior work has integrated retrieval-based exercises into second language learning systems to enhance vocabulary acquisition, reading comprehension, and communicative fluency [23, 31, 36].

When combined with microlearning [42]—an approach that delivers educational content in small, easily digestible units—retrieval practice becomes particularly effective for busy adults. Microlearning helps sustain motivation and engagement while minimizing the time burden, making it well-suited for learners balancing work and study [58, 61]. Many popular mobile-assisted language learning (MALL) systems, such as Duolingo [33], Anki [2], and Quizlet [88], leverage this principle by offering interactive exercises (*e.g.*, flashcards, multiple-choice questions, and fill-in-the-blanks). In addition, the field of HCI has explored retrieval-based microlearning in various contexts, such as spaced practice for vocabulary acquisition, adaptive exercise scheduling, and bite-sized practice integrated into daily routines [31, 36].

Building on this, our work provides a mobile practice environment that leverages retrieval practice for information workers learning practical English skills. We focus on a particular exercise type—multiple choice fill-in-the-blank questions—which are widely used in standardized proficiency tests [38]. Furthermore, recent research has demonstrated that AI-generated multiple-choice questions can reach expert-level quality [30, 39], highlighting their potential as a scalable and effective method for creating adaptive practice. This allows us to focus on the content's quality, which we ensure through our systematic generation pipeline (see Section 4.3).

3 Formative Study

To inform the design of LingoQ, we conducted an online survey with 49 information workers whose native language is Korean, followed by semi-structured interviews with ten volunteers. We aimed to understand the type of barriers they face during daily English-related tasks, limitations with the existing digital tools they use to handle these tasks, and effective EFL learning practices they have experience using to develop work-related English skills.

3.1 Procedure and Analysis

Online Surveys. Through both closed and open questions, we asked participants about the challenges they face as non-native English speakers at work, the digital tools they use to support English-related tasks, the effectiveness of their EFL learning strategies, and their perceived need for continued learning. We also asked about the willingness to participate in a follow-up interview. The online survey was advertised to native Korean speakers on social media and our internal network, inviting information workers who use computers for their work and regularly perform tasks that require English. Forty-nine people (25 females; aged 22–49) completed the survey, which included 22 researchers, 11 engineers, and 16 professionals from various fields, including strategic planning, sales and marketing, design, general affairs, and healthcare. The survey took approximately 20 minutes to complete. We compensated 5,000 KRW (approx. 4 USD) for survey respondents.

Interviews. For in-depth analysis, we conducted follow-up interviews with ten survey respondents who indicated their willingness to attend as part of the online survey. Each interview lasted about 40 minutes and was conducted in person or remotely, depending on the participants' availability. We revisited the interviewees' survey responses and asked them to elaborate on their open-ended answers. Using screen sharing and think-aloud protocols [20], participants walked through recent scenarios involving English-related tasks, demonstrating queries they had made to generative AI (*e.g.*, ChatGPT, Gemini) or other tools. They also described their review practices focused on work-specific English content. We compensated 20,000 KRW (approx. 14 USD) for interview participants.

Analysis. All interviews were audio-recorded and transcribed for analysis. We summarized the closed-ended survey questions using descriptive statistics. We used Thematic Analysis [14] to qualitatively analyze both the open-ended questions of the survey and interview transcripts. One researcher coded survey responses as well as interview transcripts simultaneously, grouping them into broader themes. The research team iterated through several rounds of discussion to refine these themes. In the following sections, we present findings from both the survey and the interviews, referring to each interview participant as I1 through I10.

3.2 Finding 1: Understanding the Difficulties of EFL Workers in the Workplace

Participants worked with large amounts of information written in English as part of their daily tasks, ranging from (1) communication through emails or messengers, (2) accessing online resources, and (3) writing professional documents such as reports or papers, most of which required intensive reading and writing rather than spoken communication.

One common linguistic difficulty that the majority of respondents (25/49) pointed out was **lexical disruption**, noting that unfamiliar domain-specific terminology often hindered their comprehension and prompted them to look up words frequently. I5, who works in governance administration at an international research lab, remarked that "*The official materials from the UN Headquarters are often overly formal and full of UN-specific terms, which slows me down as I have to look them up.*" Similarly, I1—an international business development manager—noted ,"*For example, I used to think 'airway' only meant a flight route, but later learned it also refers to a respiratory tract [a human body part].*" Relatedly, participants reported not only linguistic challenges but also affective challenges—stemming from a lack of confidence, which consequently hindered their workflow. Four participants mentioned in their surveys and interviews that they proofread emails for grammar, formality, and tone before sending, concerned that mistakes might appear impolite or give a negative impression of their professional competence.

Participants often felt that current EFL learning was **disconnected from work context**. More than 2/3 of the survey respondents (33/49) rated that they often (45.0%) or always (22.5%) feel the need for learning English for work on a 5-point Likert-type scale question. Yet, the majority of them (28/49) were not currently studying English, demonstrating the difficulty of constant engagement in EFL learning practices during work. For those who were currently studying English, all of them (21/21) reported that they used easily accessible and self-directed mobile apps, outside of work context (*e.g.*, Duolingo [33], Speak app [99]). Other common practices included online tutoring (*e.g.*, Ringle [90]; 8/21), reading English novels or articles (8/21), shadowing (4/21), and in-person courses (3/21). However, the majority of them (14/21) struggled to sustain their learning because irrelevant learning materials did not translate into practical support in their work contexts. In the follow-up interview, I10 noted, "*What I really want to learn right now is material I can use immediately in business meetings, but finding a suitable platform or tool has been very difficult.*" I1 also remarked, "*I'm often exposed to highly specialized medical terms, but when the material is from a learning app or an article outside my field, it tends to use more general vocabulary and expressions. While this is helpful for conversations, it's not very useful when reading work-related articles or clinical papers.*"

Challenges occurred during the reviewing phase as well due to **lack of sustainable review routines**. To align EFL learning with their work contexts, six interviewees once tried to review unfamiliar words and expressions from work by compiling personal glossaries and organizing them with tools such as Notion, Google Docs, or the open-source flashcard app Anki [2]. However, participants failed to maintain engagement with such review routines, as manually collecting work-related vocabulary or expressions was time-consuming and burdensome. Moreover, reviewing these materials with explicit exercises further discouraged continued practice. In the follow-up interviews, I9 noted, "*After work, I don't want to revisit the traces of what I did during the day. Reviewing would mean opening my daily logs in a workspace like Notion, finding the target words, gathering them on another page, and then asking GPT or searching Google for their meanings. Most days, it just feels too much.*" Also, I10 remarked, "*In one-on-one business English tutoring, my teacher listed my mistakes in Google Docs, but reviewing them felt like just reading meeting minutes and was neither fun nor motivating. I wish I could review them in more engaging ways, like quizzes or other formats for sustainable practice.*" While interactive apps such as Anki, with flashcard and quiz features, were available, participants (I8, I9) found it overly complex and overwhelming to customize and manually upload word lists, especially after work.

3.3 Finding 2: Common Patterns of English Language Queries

To address language barriers, all participants used language assistance tools for lookup and double-checking, including dictionaries, web search engines, translators, AI-based writing assistants (*e.g.*, Grammarly [47], DeepL [26]), and LLM-based chatbots (*e.g.*, ChatGPT [84], Gemini [46], Claude [4]). In particular, most survey respondents (46/49) commonly used LLM-based chatbots, which offered convenient conversational support and context-aware explanations for a wide range of English-related difficulties. Interviewees entered queries primarily by copying and pasting text, ranging from single words to full passages, along with a recurring prompt for linguistic support. From their usage scenarios, we identified three prominent query patterns to LLM-based chatbots: **look-up**, **translation**, and **proofreading**.

Most interviewees (7/10) often used chatbots just like dictionaries to **look up** definitions of unfamiliar words or description of confusing grammar during their tasks. I5 noted, "*These days I just ask ChatGPT when I'm unsure about grammar, like 'an MBA or a MBA?'*" I6 and I8 found LLMs useful for clarifying domain-specific terms or subtle nuances, as they offered context-specific explanations, especially for words with multiple possible meanings.

All interviewees (10/10) used chatbots to **translate** text in English to Korean and vice versa. They translated text in English into Korean to ensure their comprehension and translated Korean into English to compose professional writing more efficiently. I6 noted, "*I usually ask LLMs to align the original and the translation side by side, so I can double-check whether each part conveys the intended meaning,*" highlighting the need for a dual-language view to enable rapid comparison under time pressure at work.

The majority of interviewees (6/10) also frequently asked chatbots to **proofread** their own draft, ranging from formal business documents to casual conversation with colleagues. Participants refined grammar, tone, and style to fit the context of the communication, often by providing additional information (e.g., their relationship with the interlocutor) to the assistant. For example, I5 copied entire email threads into the LLMs and asked, "*Please proofread my reply,*" to ensure their response was both grammatically sound and aligned with the ongoing exchange. I9 even checked short messages for online meetings, such as "Will you be joining soon?", to understand the nuance of the message that they may be implying in the message: "*I worry it might sound like I'm pushing, so I ask ChatGPT to review even simple texts before sending.*"

4 LingoQ

Our formative study revealed that EFL workers suffer from *lexical disruption* while handling information work in English. In addition, our participants noted that most existing EFL learning systems rely on generic materials disconnected from their work contexts—despite research in EFL education showing that authentic, usage-based practice fosters learner engagement and improves both proficiency and self-efficacy [9, 10, 43]. To address these challenges, we designed and developed LingoQ, a language querying and self-directed learning system that provides work-related quizzes generated from language queries. In this section, we discuss our design rationales from the formative study and literature. We then describe our system design and generative pipelines, along with implementation details.

4.1 Design Rationales

DR1. Leverage AI-Assisted Language Queries as a Source of Learning Material. In our formative study, nearly all participants relied on LLM-based AI assistants for work-related English tasks, such as looking up unfamiliar terminology, resolving confusing grammar, translating text, or proofreading their writing. We therefore treated users' language queries with an AI chatbot as an authentic source from which we can learn the English assistance that they need and generate learning materials. Moreover, certain types of user queries, such as searching for a definition of a word or comparing input text with edited text, explicitly reveal their weakness in English proficiency.

DR2. Optimize AI Assistant Interface for Language Querying. Since our formative study participants frequently used generative AI assistants, such as ChatGPT, for language practice, we observed that their querying interactions were often inefficient and tedious. For example, participants had to repeatedly type boilerplate commands in the input (*e.g.*, "`Translate this into Korean:`") whenever they initiated a new query. In addition, participants often issued follow-up requests to format responses for their language tasks (e.g., requesting to display Korean and English text side by side or highlighting edited portions to track changes), which led to unnecessary back-and-forth dialogue turns.

Hence, we incorporated LingoQuery, an LLM-based assistant dedicated to language queries. Leveraging the design of typical AI assistant chatbots like ChatGPT [84] or Gemini [46], which information workers are already familiar with, we adopted a similar structure while optimizing the interactions and interfaces for such usage. Given that participants in our formative study often copied text to query digital tools, we implemented a keyboard shortcut that directly copies and pastes selected text from the computer into a new chat message. We also introduced *query intents* that users can attach to an input message, which automatically insert predefined yet customizable prompts for three frequent request types—look up, translate, and proofread—thereby avoiding manual typing of boilerplate instructions. The AI responses to these query intents are rendered in language-relevant message displays (see Figure 2).

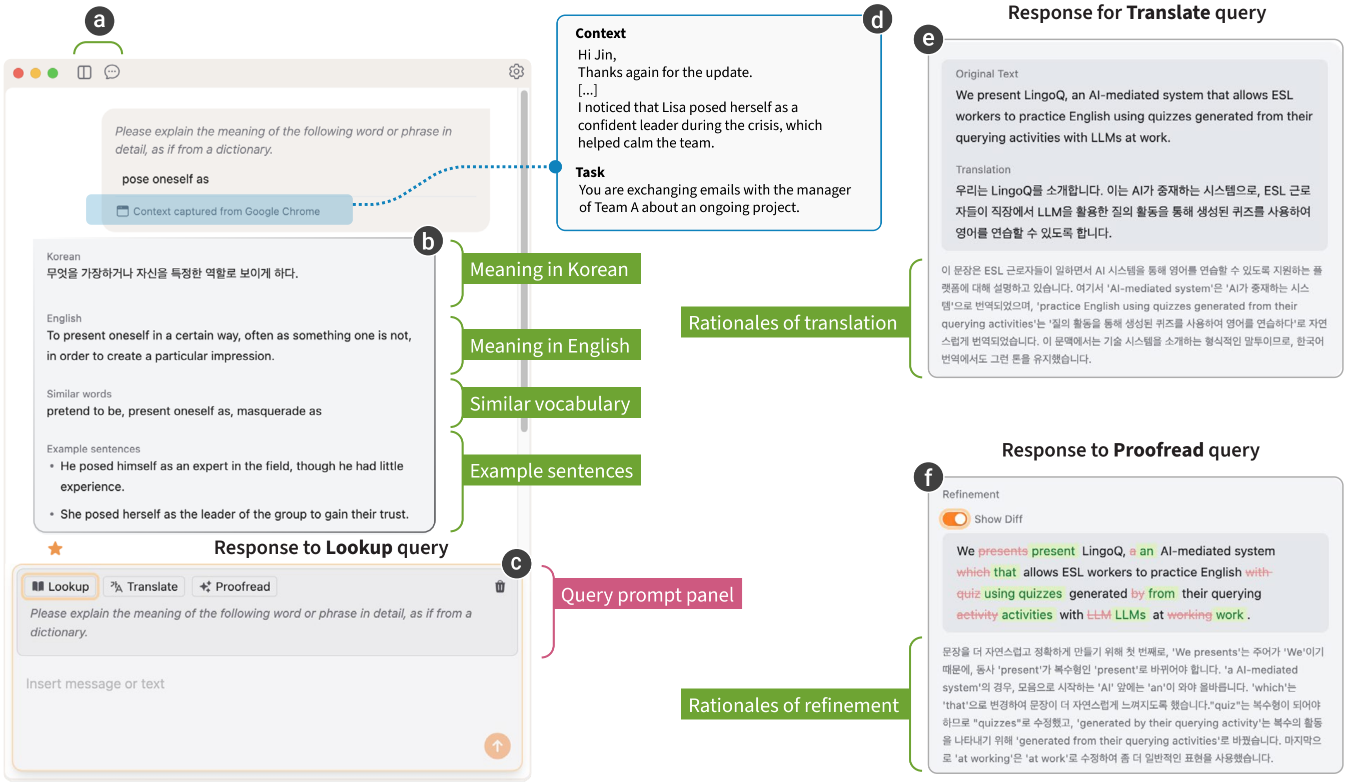
DR3. Streamline Reviewing Work-Related Language Activity. Participants in our formative study attempted to review vocabulary or expressions used in daily tasks at work, but the burden of manually collecting and revisiting materials without explicit exercises hindered sustained engagement, particularly after work. Inspired by literature suggesting that microlearning with short practice sessions embedded into daily routines can foster sustained engagement and improve proficiency [42, 58, 61], we designed our system to generate bite-sized interactive quizzes. These quizzes are generated directly from automatically collected queries, supporting continuous practice without extra burden outside work. We adapted multiple-choice fill-in-the-blank question formats from standardized tests (*e.g.*, TOEFL, TOEIC, GRE) to support vocabulary and grammar practice. While the format of problems can be diversified (*e.g.*, reading/listening comprehension, open-ended questions, short writing tasks), we limit our study material to the fill-in-the-blank multiple-choice question (MCQ) format for its simplicity to support easy access to English study anytime, anywhere on a mobile phone. The effectiveness of other question formats lies beyond the focus of this work, which is to generate study material relevant to work.

Based on these design rationales, we developed LingoQ that consists of LingoQuery (4.2), LingoQuiz (4.4), and the backend pipeline (4.3). LingoQuery is a desktop-based AI assistant that users can share English-related queries or discuss freely. The backend pipeline manages user query data from LingoQuery and generates questions and curates them into quizzes. LingoQuiz is a mobile application that offers 10-question quizzes generated from the user's dialogues with LingoQuery.

4.2 LingoQuery

4.2.1 Interaction Components of LingoQuery. LingoQuery adopts the typical interface design of desktop versions of LLM-based AI assistants, such as ChatGPT [84] and Claude [4], while incorporating bespoke interaction components tailored to the English-language query contexts. The sequence of chat messages is organized into chat threads, and users can either start a new thread or append messages to existing ones by selecting them from the sidebar (ⓐ in Figure 2). By default, the AI response messages are rendered as a markdown-formatted view.

Language Query Intent Selection and Prediction. The system supports three predefined query intents: (1) *Look up*, (2) *Translate*, and (3) *Proofread*. When composing a new message, users can explicitly select a query intent in the chatbox by pressing the buttons at the top, which load a predefined prompt (ⓒ in Figure 2). Each query intent applies a prompt template that is concatenated with the user's message input; for example, "`Please explain the meaning of the following word (or expression) in detail in dictionary format`" for Look up, which the user can edit before sending. If no query intent is selected, the message in the chatbox is treated as a plain prompt in text, and the system automatically predicts its query intent when generating a response. If a user's query corresponds to one of the three query intents, the system offers a customized view that highlights the structured information of the intent type. The *Look up* response type follows a typical dictionary format (ⓑ in Figure 2); the *Translate* response



Please explain the meaning of the following word or phrase in detail, as if from a dictionary.

pose oneself as

Context captured from Google Chrome

Korean
무엇을 가장하거나 자신을 특정한 역할로 보이게 하다.

English
To present oneself in a certain way, often as something one is not, in order to create a particular impression.

Similar words
pretend to be, present oneself as, masquerade as

Example sentences
- He posed himself as an expert in the field, though he had little experience.
- She posed herself as the leader of the group to gain their trust.

Please explain the meaning of the following word or phrase in detail, as if from a dictionary.

Insert message or text

Context
Hi Jin,
Thanks again for the update.
[...]
I noticed that Lisa posed herself as a confident leader during the crisis, which helped calm the team.

Task
You are exchanging emails with the manager of Team A about an ongoing project.

Original Text
We present LingoQ, an AI-mediated system that allows ESL workers to practice English using quizzes generated from their querying activities with LLMs at work.

Translation
우리는 LingoQ를 소개합니다. 이는 AI가 중재하는 시스템으로, ESL 근로자들이 직장에서 LLM을 활용한 질의 활동을 통해 생성된 퀴즈를 사용하여 영어를 연습할 수 있도록 합니다.

이 문장은 ESL 근로자들이 일하면서 AI 시스템을 통해 영어를 연습할 수 있도록 지원하는 플랫폼에 대해 설명하고 있습니다. 여기서 'AI-mediated system'은 'AI가 중재하는 시스템'으로 번역되었으며, 'practice English using quizzes generated from their querying activities'는 '질의 활동을 통해 생성된 퀴즈를 사용하여 영어를 연습하다'로 자연스럽게 번역되었습니다. 이 문맥에서는 기술 시스템을 소개하는 형식적인 말투이므로, 한국어 번역에서도 그런 톤을 유지했습니다.

Refinement

We ~~presents~~ present LingoQ, ~~a~~ an AI-mediated system ~~which~~ that allows ESL workers to practice English ~~with quiz~~ using quizzes generated ~~by~~ from their querying ~~activity~~ activities with ~~LLM~~ LLMs at ~~working~~ work .

문장을 더 자연스럽고 정확하게 만들기 위해 첫 번째로, 'We presents'는 주어가 'We'이기 때문에, 동사 'present'가 복수형인 'present'로 바뀌어야 합니다. 'a AI-mediated system'의 경우, 모음으로 시작하는 'AI' 앞에는 'an'이 와야 올바릅니다. 'which'는 'that'으로 변경하여 문장이 더 자연스럽게 느껴지도록 했습니다."quiz"는 복수형이 되어야 하므로 "quizzes"로 수정했고, 'generated by their querying activity'는 복수의 활동을 나타내기 위해 'generated from their querying activities'로 바꿨습니다. 마지막으로 'at working'은 'at work'로 수정하여 좀 더 일반적인 표현을 사용했습니다.

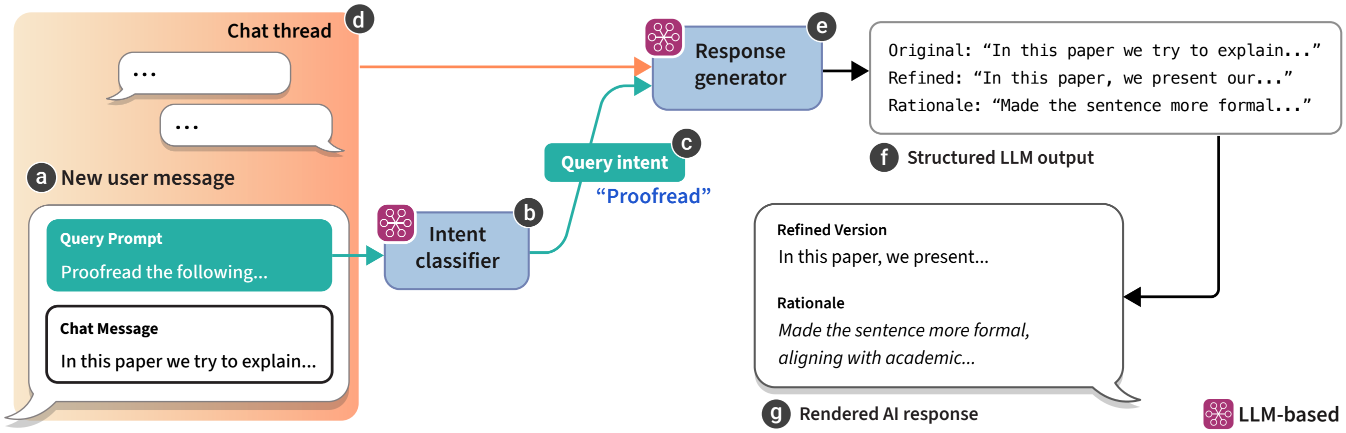
Figure 2: Main window and the interface components of LingoQuery. Users can open new chat threads of their choice in the thread list or via the New Chat button (ⓐ). The AI's responses for three major types of query intents—*Look up* (ⓑ), *Translate* (ⓔ), and *Proofread* (ⓕ)—provide the UI components tailored to each query type. The new message panel (ⓒ) incorporates a query prompt panel at the top. Users can insert a template query prompt in their message via the three quick-access buttons. When a query is sent via a keyboard shortcut ⌨, the system analyzes a screenshot of the user's active window, and the user can review both the surrounding context of the copied sentence and the task at the time of the query ⓓ. By clicking the star icon ★ below AI responses (ⓑ), users can mark the message to increase the likelihood that the corresponding question will be included in LingoQuiz.

type provides a side-by-side view for comparing original and translated text (ⓔ in Figure 2); and the Proofread response type displays a formatted container showing the proofread text with the rationale for edits underneath, along with an option to toggle track changes so users can quickly see where edits were made (ⓕ in Figure 2). These views were informed by how interviewees in the formative study customized responses when using generic LLM-based assistants.

Shortcuts and Contexts. To enable users to receive assistance within the context of work-related applications, LingoQuery provides an operating system–level shortcut to trigger a query. When the user highlights text anywhere on the computer and presses '`Ctrl + Cmd + C`' on MacOS or `Ctrl + Alt + C` on Windows, the LingoQuery window opens with a new chat thread and the copied text pre-filled in the chatbox. At the same time, the system captures a screenshot of the active window and displays it alongside the text when the shortcut is pressed. Before submitting the query, the user can choose to include the screenshot with the message or remove it if it contains sensitive information. If the screenshot is included, the system runs image understanding on OpenAI's `GPT-4o` model to extract the text surrounding the copied content and infer the nature of the tasks based on metadata of the visible application on a screen, enriching the content for question generation later. The inferred context will be displayed in the UI as well.

Marking Messages for Prioritizing Question Generation. Users can also mark noteworthy AI responses via the ★ star icon (right above ⓒ in Figure 2). Marked messages are processed in the system pipelines, increasing the likelihood that the corresponding question will be included in a quiz. This mark feature allows users to flag particular words or expressions they wish to review within the context of LingoQuery, removing the need to manually track what they want to study.

4.2.2 LingoQuery Conversational Pipelines. LingoQuery is a self-contained app that generates responses to user requests, similar to an LLM-based AI assistant with additional customization for specific query types. A user's query input is processed before being sent to the LLM engine (OpenAI API in our case). Figure 3 illustrates



```
Original: "In this paper we try to explain..."
Refined: "In this paper, we present our..."
Rationale: "Made the sentence more formal..."
```

Refined Version
In this paper, we present...

Rationale
Made the sentence more formal, aligning with academic...

Figure 3: Conversational pipeline of LingoQuery. When the user sends a new message ⓐ, the intent classifier ⓑ identifies the query intent ⓒ, which is then passed to the response generator ⓔ together with the chat history ⓓ. The response generator produces an appropriate response ⓕ structured according to the query intent. Finally, LingoQuery renders this structured response accordingly ⓖ.

the response generation pipeline of the LingoQuery conversational agent when it receives a new user message (ⓐ in Figure 3). LLM-based **Intent Classifier** (ⓑ in Figure 3; see Section A.1 for the instruction provided to the LLM) determines the corresponding query intent (ⓒ in Figure 3). Both the chat history (ⓓ in Figure 3) and the detected intent are then passed to the **Response Generator** (ⓔ in Figure 3), which produces an AI response using an LLM (See Section A.2 for the instruction provided to the LLM). When the query intent does not correspond to a plain-text message but instead falls into one of the intents Look up, Translate, or Proofread, the LLM output is returned as a JSON object containing relevant attributes (*e.g.*, the *original* input, *refined* text, and the *rationale* of refinement in the case of a Proofread intent). The structured format enables the application interface to render the output through a bespoke UI (ⓖ in Figure 3).

4.3 Question Generation Pipelines

The question generation in LingoQ follows three pipelines—question generation, question quality evaluation for filtering, and question selection—that transform interactions collected through LingoQuery into validated quiz questions for LingoQuiz.

4.3.1 Generating Questions from Conversation and Context. Figure 4 illustrates the question generation pipeline, which periodically produces fill-in-the-blank multiple-choice questions from query-response pairs collected in LingoQuery and captured context data. Every five minutes, the system checks for new query–response pairs collected from LingoQuery (ⓐ in Figure 4). For each pair, an LLM-based module evaluates whether the query is English-related (ⓑ in Figure 4); if not, it is filtered out from the question generation pipeline. For eligible queries, the system takes the eligible pair and additional information—the conversation history with contextual data captured from the screenshot—to initiate generation (ⓓ in Figure 4). The pipeline applies a different system prompt with few-shot examples modeled after TOEFL, TOEIC, and GRE example questions (ⓔ in Figure 4). For each conversation, two distinct questions are generated to ensure variety (ⓕ in Figure 4). Each question will generate one structured output in JSON format that contains: stem with a blank, key, distractors, explanation, and rationale for question generation (ⓖ in Figure 4).

Finally, the contextual information extracted from a screenshot and conversation history is fed into the question generation pipeline to produce questions aligned with that context. The question generation module ensures that the question stem is relevant to the worker's task context. For example, a user query may simply involve searching for a word (e.g., "airway"), in which case the surrounding text (e.g., "the patient's airway to ensure proper breathing") from the screenshot can be used to extract the context and generate a relevant question stem. When a user submits an answer in LingoQuiz, the explanation provided will include this context to supplement the rationale for the correct answer.

4.3.2 Quality Assurance of Generated Questions. To ensure the quality of generated questions, they are evaluated by an LLM-based evaluation module informed by prior literature [30, 39]. The evaluation applies two binary criteria: answerability, that is, whether the question can be clearly and correctly answered, and proficiency, that is, whether the question requires an appropriate level of English skill and is not too easy to answer (ⓗ in Figure 4). Questions that fail one of the criteria are iteratively refined—up to three times in total—by feeding the evaluator's rationale for failure, along with the original input, back into the question generation module (ⓘ in Figure 4). If a question passes within three iterations, it is added to the question pool (ⓙ in Figure 4), otherwise discarded.

4.3.3 Question Pool and Selection Logic. After quality checks, questions undergo a final format validation to ensure that all required components—stem with blank, distractors, key, and explanation—are properly structured. Validated questions are then added to the question pool. When a user initiates a quiz, ten questions are drawn from the pool. Each quiz contains 10 questions: 7 are selected from newly

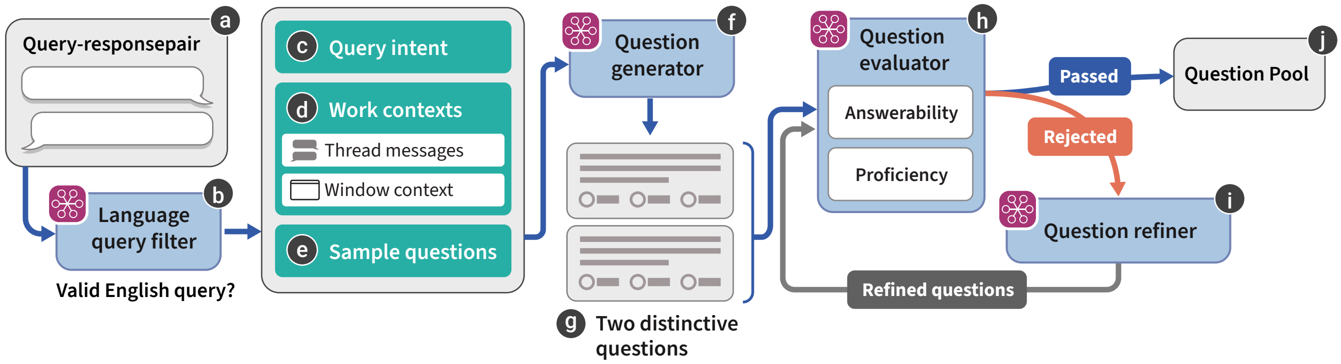


Figure 4: Question generation pipeline of LingoQ. When a query–response pair arrives ⓐ, the language query filter ⓑ identifies the query intent ⓒ, which is then passed to the Question generator ⓕ together with work contexts ⓓ and exam samples ⓔ. The generator produces two candidate questions ⓖ, which are evaluated by the Question evaluator ⓗ on two criteria: answerability and proficiency. The Question refiner ⓘ refines each question up to two iterations, and items that still fail are discarded. Accepted questions are stored in the question pool ⓙ.

generated questions, and the remaining 3 are randomly drawn from the pool of questions that a user has solved previously using weighted probability. The system assigns higher weights to questions that have been repeated less frequently, answered incorrectly in the past, marked with the ★ star icon in LingoQuery, or not practiced recently. As a result, each quiz balances new questions (70%) with questions that workers need to revisit. The examples of generated questions for each type are available in Figure 6.

4.4 LingoQuiz

Users can practice work-related English vocabulary and grammar by solving questions in LingoQuiz, generated from their dialogues in LingoQuery. LingoQuiz provides a dashboard (A in Figure 5) that helps users track how many quizzes they have completed that day, how many they have completed in total since the beginning, and how many new questions are available in the question pool (① in Figure 5). Clicking 'Start Quiz' launches 10 multiple-choice questions, each requiring users to fill a blank (B in Figure 5). When the question is generated from a marked AI response or when it is the first attempt, it is displayed with a ★ or 'new!' badge (② in Figure 5). Questions generated from screenshots or sufficient thread context include the inferred task as a hint (③ in Figure 5). When users press the Submit button after selecting an option, the app provides immediate feedback on whether the answer is correct and the explanation (④ in Figure 5). Within the quiz, any incorrect questions reappear until users provide the correct answer. Once all ten questions are answered correctly, the progress bar completes (⑤ in Figure 5), and the quiz ends with a completion screen (C in Figure 5). Afterward, users may proceed to a new quiz or return to the dashboard (A in Figure 5).

4.5 Implementation

We implemented the core system in Python running on a FastAPI [41] server that provides REST APIs for both LingoQuiz and LingoQuery. The chat history, generated quizzes, and user interaction data are stored in a PostgreSQL [49] database on the server. The conversation and the question generation pipelines leverage OpenAI's Chat Completion APIs [85] on top of the LangChain [67] framework to run the underlying LLM inferences. All LLM inference and image understanding tasks are performed using a `gpt-4o` model. To protect user queries that may contain sensitive information, we used OpenAI Enterprise, which neither uses our data for training nor retains them.

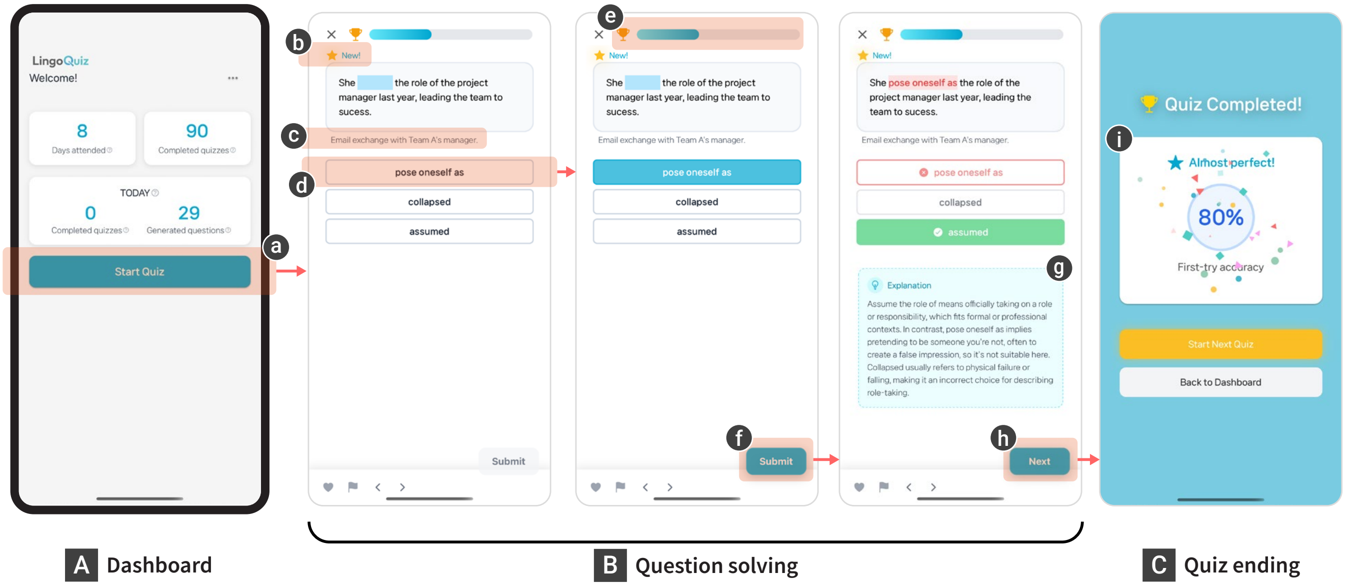
We built LingoQuery as a cross-platform desktop application using Electron [86], to support both Windows and MacOS desktop computers. The LingoQuiz app was implemented using React Native [78] as a cross-platform mobile application running on both iOS and Android phones. Both apps were written in TypeScript [79] and communicate with the server via REST API.

5 Deployment Study

We conducted a three-week field deployment study with 28 EFL workers. To address our research questions, we aimed to examine how EFL workers engage with LingoQ and how it affects their English proficiency and self-efficacy. In addition, we conducted an expert evaluation to assess the quality of questions generated by LingoQ. The study was conducted in South Korea with Korean native speakers and approved by the Institutional Review Board.

5.1 Participants

We conducted power analysis to calculate the number of participants necessary. With an expected medium effect size of 0.5, the required sample sizes are 27 for a paired t-test (α = .05, one-tailed, power = .80) and 28 for a Mann–Whitney test (α = .05, one-tailed, power = .80). Our inclusion criteria were information workers who are: (1) working at least 30 hours per week, (2) using a computer as the primary work tool, (3) regularly performing tasks involving English, such as information access, communication, and document writing, (4) being a native Korean speaker, and (5) being an EFL learner. We advertised our study on social media. Initially, we



She ___ the role of the project manager last year, leading the team to sucess.

Email exchange with Team A's manager.

She pose oneself as the role of the project manager last year, leading the team to sucess.

Explanation

Assume the role of means officially taking on a role or responsibility, which fits formal or professional contexts. In contrast, pose oneself as implies pretending to be someone you're not, often to create a false impression, so it's not suitable here. Collapsed usually refers to physical failure or falling, making it an incorrect choice for describing role-taking.

Figure 5: Main screens of LingoQuiz. In the Dashboard screen A, users can check their records and stats, along with the number of new questions add to their question pool today. When starting a quiz by pressing the Start Quiz button ⓐ, a quiz with 10 unique questions are provided sequentially B. The new question that appears to the user for the first time is indicated by the star icon (ⓑ). For questions generated from messages with context, the task description is provided (ⓒ). To solve the question, the user can select an option (ⓓ) and press the Submit button (ⓕ) to submit an answer. Then the result the question is shown immediately, with explanation (ⓖ), regardless of whether the user had selected a correct answer or not. After solving the ten questions, questions with wrong answers appears again, until all are answered correctly. The progress bar (ⓔ) indicates the current progress. In the Ending screen C, users can practice a new quiz or return to the Dashboard screen.

recruited 34 workers to account for potential attrition; two participants dropped out due to their corporate security policies that prevented them from installing LingoQ, and four were later excluded during analysis for not meeting the minimum requirements. Finally, a total of 28 EFL workers (Table 1; P1–P28; 18 females and 10 males) completed the study.

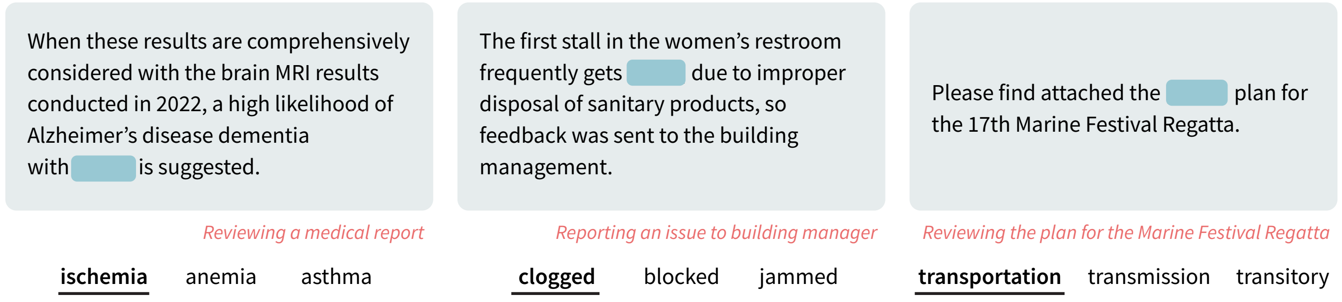
Participants were aged between 25 and 48 years old (M = 33.4, SD = 6.5) and represented diverse professional domains. Based on CEFR [82][2] self-assessed English proficieny, 3 participants identified themselves as *A1* (beginner), 4 as *A2* (elementary), 14 as *B1* (intermediate), and 7 as *C1* (advanced). All participants reported using LLM-based chatbots daily during their workdays. Additionally, they had prior experience with EFL learning for work, including tutoring (17), vocabulary apps (13), and English media (12). As a minimum requirement for study completion, we instructed participants to use LingoQuery for at least 10 days during the three-week period and LingoQuiz for at least 10 days during the same period. To qualify as having used an app on a given day, participants needed to submit at least two questions in LingoQuery and complete at least one quiz in LingoQuiz. As compensation for their participation, we offered 200,000 KRW (approx. 144 USD) based on the required system usage over three weeks.

5.2 Procedure

Pre-Study Preparation. Upon sign-up, we sent participants a link to a pre-study survey and a pre-study English proficiency test. The survey included three items on a 5-point Likert scale that assessed the perceived relevance, effectiveness, and engagement of their past EFL learning methods, along with 16 items from the Questionnaire of English Self-Efficacy (QESE; eight on reading and eight on writing) on a 7-point scale, excluding speaking and listening to align with our research focus [106].

The English proficiency test consisted of 28 multiple-choice items selected from TOEIC (Test of English for International Communication) [38], a standardized English proficiency test for general business. We did not include spoken English proficiency measures because they lie beyond the focus of this work. The test included two types of questions: 16 simple fill-in-the-blank items—each with a single sentence and one blank—and three sets of four fill-in-the-blank items, each requiring participants to complete blanks within a single paragraph. Before the study, we finalized the items from 46 questions by administering them to 29 information workers—who are not our study participants—and selecting those whose percentage of correct answers fell between 40% and 80%, as suggested in Classical Test Theory [27]. (see Appendix B for details of the item validation.)

[2] The Common European Framework of Reference for Languages (CEFR) defines proficiency levels as basic (A1: beginner; A2: elementary), independent (B1: intermediate; B2: upper-intermediate), and proficient (C1: advanced; C2: proficient).



When these results are comprehensively considered with the brain MRI results conducted in 2022, a high likelihood of Alzheimer's disease dementia with ___ is suggested.

Reviewing a medical report

<u>**ischemia**</u> anemia asthma

(a) P15

The first stall in the women's restroom frequently gets ___ due to improper disposal of sanitary products, so feedback was sent to the building management.

Reporting an issue to building manager

<u>**clogged**</u> blocked jammed

(b) P20

Please find attached the ___ plan for the 17th Marine Festival Regatta.

Reviewing the plan for the Marine Festival Regatta

<u>**transportation**</u> transmission transitory

(c) P11

Figure 6: Selected questions actually generated by LingoQ during the deployment study for three participants. Each question consists of a stem with a blank, context reminding the task at the time of query (red text), and three alternatives with correct answers underlined.

Onboarding Session. A group of 2-3 participants attended a 1-hour, in-person onboarding session, bringing their laptop to install LingoQuery. After explaining our study goals, we assisted participants with installing the system on their laptops and mobile devices. To ensure participants fully understood how to use the system, we conducted a hands-on tutorial that covered the main features of LingoQ. This step ensured that participants had quizzes available during the early stage of the deployment period.

Deployment. Immediately after the onboarding session, participants began using LingoQ for three weeks. During this period, participants were instructed to direct their English-related queries to LingoQuery, instead of using ChatGPT or other tools. They were instructed to use LingoQuiz at any time of the day.

At the end of each week, we sent participants a message summarizing LingoQ usage to remind them of the minimum requirement for study completion. Additionally, when participants were inactive for more than three days, LingoQuiz sent an evening push notification. LingoQuiz notified participants if new questions were available for a day.

Post-study Survey. After the 3-week deployment period, we sent an online post-study survey and a post-study English proficiency test. The survey reassessed QESE [106] and the three ratings of perceived learning experience used in the pre-study survey, but targeted for LingoQ, and was supplemented with follow-up questions probing the reasons for participants' ratings. It also asked participants about their willingness to use LingoQ on a 5-point Likert scale and included open-ended questions regarding their overall user experiences and suggestions for design improvements. For the post-study English proficiency test, we used the same question set as in the pre-study test, with both the question order and the answer-option order randomized to mitigate test–retest bias. Participants were not shown the correct answers after the pre-study test, ensuring that they could not learn directly from the test itself. Using identical questions is a common method for controlling variation in question-set difficulty (*c.f.*, [52, 58, 64, 71, 107]). The three-week deployment period between the pre- and post-tests also provided a sufficiently long interval to minimize memory and practice effects [35].

5.3 Expert Evaluation of Questions

To assess the performance of the question generation pipeline (Section 4.3.2), we conducted an expert evaluation using a subset of questions generated during the deployment study. We randomly sampled 30 questions: 24 in the question pool (*i.e.*, those which passed the quality checking) and 6 that were eventually discarded due to unmet answerability or proficiency criteria. 17 of 24 questions in the pool had been presented to users during the study. To enable a direct comparison with the question generation pipeline, we applied the same criteria used by LingoQ's Question Evaluator (*c.f.*, Figure 4-ⓗ)—answerability and proficiency. Informed by prior work on educational question quality evaluation [30, 39], we developed a rubric comprising three items: one directly aligned with answerability and two aligned with proficiency. These items collectively operationalize our two evaluation criteria (see Appendix C for the full rubric and mappings).

We recruited three English educators (E1–E3; all female) through social media advertisements. Their professional backgrounds included university and high school teaching, as well as the development of standardized English test. They were aged 39, 53, and 39, with 10, 20, and 15 years of experience in English education, respectively. The experts participated in remote evaluation sessions via Zoom. In the session, experts first evaluated the 30 questions through an online survey. They then went through a follow-up interview, where we asked about the potential benefits and concerns of the approach we take in LingoQ, specifically generating EFL learning materials from their work context. The evaluation and interview took 90 minutes to complete. We compensated them with 100,000 KRW (approx. 72 USD).

5.4 Data Analysis

To examine participants' engagement and usage patterns with LingoQ, we conducted descriptive analyses of interactions with both LingoQuery and LingoQuiz. For LingoQuery, we analyzed message–response pairs, usage days and patterns, and the distribution of query prompts. For LingoQuiz, we examined the number of solved questions, usage days, and solving patterns, and quiz progress across repeated attempts.

Table 1: Demographic information and self-reported CEFR levels of the participants in our deployment study.

Alias	Age	Gender	CEFR	Job title
P1	28	Female	C1	Global Business Developer
P2	28	Female	A2	Software Engineer
P3	39	Female	B2	Hospital Operations Manager
P4	28	Male	A2	Medical Resident
P5	32	Female	B1	Product Designer
P6	36	Female	B2	Software Engineer
P7	47	Female	B1	Kindergarten Counselor
P8	45	Female	B2	Administrative Coordinator
P9	37	Female	B2	Office Manager
P10	29	Male	A2	International Patient Coordinator
P11	30	Male	B2	Sports Event Manager
P12	38	Female	B2	Export–Import Specialist
P13	42	Male	C1	Professor
P14	35	Female	B2	Apparel Export Manager
P15	29	Male	A2	Clinical Psychology Trainee
P16	41	Male	A1	IT Security Manager
P17	27	Female	B1	Graduate Student
P18	28	Male	B1	Machine Learning Engineer
P19	25	Male	B2	Graduate Student
P20	29	Female	C1	Real Estate Professional
P21	32	Female	A1	Biotech Researcher
P22	48	Female	C1	Logistics Specialist
P23	25	Female	B1	Graduate Student
P24	30	Female	B2	Marketing and Project Manager
P25	33	Male	A1	Network Engineer
P26	31	Female	C1	Nuclear Policy Researcher
P27	32	Female	C1	Sports Event Manager
P28	30	Male	C1	Governance Administrator

To evaluate the generated questions, we examined pipeline performance and human evaluations. Pipeline performance was evaluated on 30 generated questions by comparing its binary judgments with expert labels. Ratings from three experts were aggregated by majority vote for answerability and proficiency, coded as true or false (with "unknown" mapped to false), and used as ground truth. We then calculated precision, recall, and F1-score, with precision measuring agreement on pipeline-accepted items, recall measuring agreement on expert-accepted items, and F1 as their harmonic mean. We transcribed the audio-recordings of the follow-up interviews with experts and conducted a qualitative analysis. One researcher coded the data using initial themes informed by the interview guide. The full research team then refined these themes through multiple rounds of peer debriefing, which surfaced the following themes: (1) the quality of AI-generated questions, (2) differences between LingoQ questions and standardized English proficiency tests, and (3) considerations in English question design.

We analyzed pre- and post-study surveys and tests to assess the effects of LingoQ on participants' learning performance and experiences. To examine changes in learning performance over the study period, we analyzed the pre- and post-study English proficiency test using a mixed-effects model and the English self-efficacy questionnaires (QESE) using paired-samples t-tests. The QESE met the normality assumption (Shapiro–Wilk test, $W = 0.95$, $p > 0.05$), allowing for parametric comparisons [51]. To investigate learning experiences relative to prior practices, we compared participants' perceived effectiveness of LingoQ with their past EFL learning using the Wilcoxon signed rank test across three criteria: relevance of materials to work, helpfulness for actual work tasks, and sustainability of engagement.

To add more nuances to the quantitative findings in describing participants' experience and perceptions, we grouped participants' answers to the open-ended questions from the post-study surveys according to the following aspects: (1) the quality and relevance of generated questions, (2) effects of LingoQ on learning and self-efficacy, and (3) the perceived benefits and drawbacks of LingoQ. We incorporate this information in different sections of findings.

6 Findings

In this section, we present findings from our deployment study in three parts. In Section 6.1, we provide an overview of the usage patterns of LingoQ and report on participants' self-report sustainability (RQ1). In Section 6.2, we investigate the quality of generated quizzes and how the use of LingoQ supported participants' English proficiency and self-efficacy (RQ2). Lastly, in Section 6.3, we report on participants' perceived utility of LingoQ and summarize their feedback on the strength and drawback of our approach (RQ3).

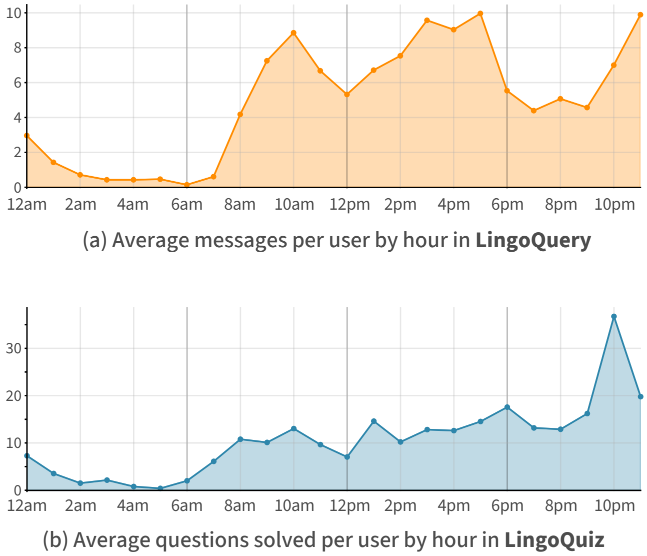
6.1 Usage Patterns and Sustained Engagement

The interaction logs and usage data indicated that participants actively engaged with LingoQ, frequently using both LingoQuery and LingoQuiz. Here we report the descriptive statistics regarding participants' usage patterns and engagement with the two apps.

6.1.1 Active Querying with LingoQuery. Across three weeks, participants opened a total of 652 conversation threads, and submitted 3,325 messages (M = 118.8 per participant) through LingoQuery. On average, participants used LingoQuery for 13.2 days (SD = 2.5, min = 10 [P15], max = 19 [P5]), exceeding the required 10 days of use. This indicates that participants engaged on most weekdays. Figure 7a presents participants' hourly engagement patterns, showing peak usage during work hours, particularly around 17 o'clock.

Participants actively used pre-defined query prompts—*i.e.*, *Look-up*, *Translate*, and *Proofread*— or wrote their own when submitting a message. Out of the 3,325 query-response pairs, *Translate* responses were the most common (1,271 reponses; 38.2%), followed by *Look up* (399 responses; 12.0%) and *Proofread* (287 responses; 8.6%). The rest of the responses (1,369 responses; 41.2%) were plain text messages, such as responses to queries asked in plain text or follow-ups.

Regarding features, eighteen Participants regularly used the *marking* feature ★, which ensures that the particular message pairs would appear in future quizzes. They marked 13.4 AI responses per person on average (SD = 16.1). Of the 241 marked messages, 91 messages (37.8%) were responses for *Look up* queries,



(a) Average messages per user by hour in **LingoQuery**

(b) Average questions solved per user by hour in **LingoQuiz**

Figure 7: Hourly engagement patterns of LingoQ during the three-week deployment. The orange line plot shows the average number of user messages in LingoQuery per hour across a 24-hour day, with notable peaks around 10 a.m., 5 p.m., and 11 p.m. The blue line plot shows the average number of solved questions per user per hour in LingoQuiz across a 24-hour day, with a clear peak around 10 p.m.

indicating participants' desire to revisit vocabulary or expressions. Participants opened LingoQuery by directly capturing the selected text and surrounding context using keyboard shortcuts ⌨, for 6.9% of all queries. However, only three participants (P11, P19, P26) dominated the usage of this feature and accounted for 65.8% of all shortcut-triggered messages.

6.1.2 Question Refinement and Validation. Of the 3,325 query-response pairs, our question generation pipeline classified 2,711 (81.5%) pairs as English language queries suitable for quiz generation, whereas 614 pairs (18.5%) were excluded because they were not English-related queries. Starting from the 2,711 English query pairs, the pipeline first generated 5,422 questions—twice the number of input pairs. Figure 8 summarizes the validation and refinement steps starting from these 5,422 initial questions (ⓐ in Figure 8). After the initial validation (ⓑ in Figure 8), 2,692 (49.7%) questions passed the evaluation. 1,114 (20.5%) questions passed on the second attempt after one refinement (ⓒ in Figure 8), and 656 (12.1%) passed after two refinement iterations (ⓓ in Figure 8). After these refinements, 960 (17.7%) questions did not satisfy the evaluation criteria and were filtered out. As a results, 4,462 validated questions were added to the question pool (ⓔ in Figure 8) over the three weeks. Of these, 3,290 were eventually exposed to participants on LingoQuiz (117.5 per participant).

6.1.3 Consistent Language Practice with LingoQuiz. Including the reappeared cases, participants solved a total of 7,155 questions (255.5 per participant). These questions were curated in 604 quizzes and participants completed most ones, leaving only 10 quizzes incomplete (1.7%) throughout the study period. Over the three weeks, Participants completed at least one quiz in LingoQuiz for 13.4 days on average (*SD* = 2.8, *min* = 10 [P27], *max* = 19 [P21]), indicating similar compliance with LingoQuery. Participants completed an average of 1.04 quizzes per day (*min* = 0.32 [P9], *max* = 2.11 [P8]), spending about 9.3 minutes (*SD* = 3.0) per quiz. This result is more than twice the number of quizzes required to qualify for study completion; the minimum requirement was 10 days out of 21, at least one quiz per day, or roughly 0.5 quizzes per day on average. Figure 7b presents participants' hourly engagement patterns, showing a more than twofold increase in usage around 10 p.m. This pattern aligns with trends observed in other popular mobile applications.

Of 3,290 unique questions, 927 questions (28.2%) appeared more than once, with the most frequently reappeared item occurring 15 times. On average, each question appeared across 2.86 quizzes (*SD* = 1.48). As the questions were repeatedly presented, participants became more likely to answer them correctly. The average accuracy for questions presented for the first time was 82.6%, which gradually increased to 89.5% upon the second exposure in another quiz, and further to 92.4% upon the third exposure.

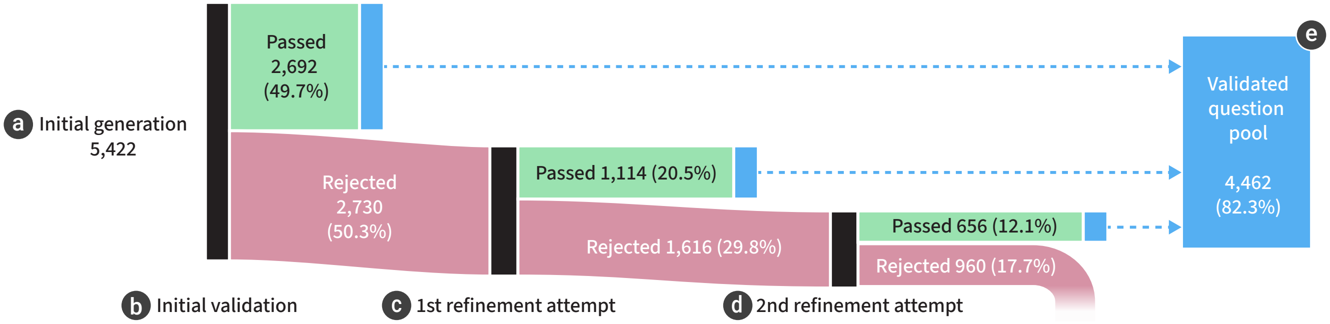


Figure 8: Overview of the question generation and validation to final question pool in the deployment study, starting from initially generated 5,422 questions ⓐ to the 4,462 validated questions ⓔ after three-stage validations with two refinements. The black vertical bars (ⓑ, ⓒ, ⓓ) denote question evaluation steps.

6.1.4 Sustained Engagement. The Wilcoxon signed-rank test showed that the participants rated the sustainability of learning with LingoQ ($M = 3.96, SD = 0.88$) significantly higher than their prior EFL learning experiences ($M = 2.96, SD = 0.79$), $z = -3.47$, $p < 0.001$ (see Figure 9a).

6.2 Learning Experience Evaluation

Over the three-week study, the use of LingoQ significantly enhanced participants' self-efficacy in using English at work. Proficiency gains varied by self-reported CEFR levels, with greater benefits for lower-level learners. These improvements were positively associated with LingoQuery usage.

6.2.1 Expert Evaluation. We compared the expert's assessments of the quality of genarated questions with the assessment from our automated quality assessment pipeline. Our comparison yielded precision/recall/F1-scores of 0.91/0.81/0.86 for Answerability and 0.85/0.92/0.88 for Proficiency, suggesting that our automated filtering provided highly aligned decisions compared with the experts' judgment, with minor discrepancies. We identified two main reasons for the discrepancies. First, experts often marked domain-specific questions as "unknown" (5 items for Answerability, 10 items for Proficiency), making it difficult to assess their quality as even the experts were not familiar with domain-specific terms. Second, they applied a higher bar for proficiency, according to the follow-up interview, as they were accustomed to carefully adjusting difficulty to learners, to maintain discrimination of the test design.

In the follow-up interviews, all experts emphasized that the key difference between LingoQ questions and standardized English tests is in the contexts used in the question stems, as E1 noted: "*TOEIC usually covers general business contexts, but some of the questions from LingoQ required knowledge confined to highly specific domains.*" Two experts (E2, E3) valued the domain-specific stems, noting that the learners' specialty in the domain could foster learning engagement and motivation, which are crucial factors for effective self-directed learning. Meanwhile, E3 remarked that the overall quality of LingoQ questions was comparable to text-completion items in TOEIC or TOEFL, noting that some items (*e.g.*, Figure 6c) resembled high-quality questions that could plausibly appear on actual standardized English tests.

6.2.2 English Proficiency. A mixed-effects model analysis revealed a significant main effect of time on English proficiency scores, with an average increase of 1 point across all participants ($p = 0.01$) (see Figure 10). Post-hoc pairwise comparisons revealed that only basic (CEFR A) learners showed a significant improvement, gaining an average of 4 points (Total 30 points) from pre-to post-test ($p = 0.01$), whereas independent (CEFR B) and proficient (CEFR C) participants showed no remarkable change. However, for the independent group, the interaction between time and the number of LingoQuery messages was significant ($p = .01$). The result indicates that more frequent use of the LingoQuery was associated with greater learning gains among these EFL workers. We further explain the learning effects of querying activity based on participants' general reactions in Section 6.3.

6.2.3 English Self-Efficacy. The paired t-test revealed significant improvement in QESE score from pre- ($M = 77.43, SD = 14.31$) to post-study ($M = 84.75, SD = 13.21$) measurements ($t(27) = -4.30$, $p < 0.001$, with 9.5% gain (see Figure 11). In addition, both the reading and writing subscales of QESE also demonstrated significant gains ($t(27) = -3.67, p < 0.01$ for reading, and $t(27) = -4.29$, $p < 0.001$ for writing), respectively. In the post-study survey, P15 highlighted the enhanced self-efficacy as the most notable benefit of LingoQ, noting "*What improved the most was my confidence. It was really satisfying to go over the mistakes I often made, and over time, I found myself reading tough sentence structures much more easily.*" Additionally, an expert from the expert evaluation reinforced this point: "*Confidence is a key factor in conversational ability, as it often*

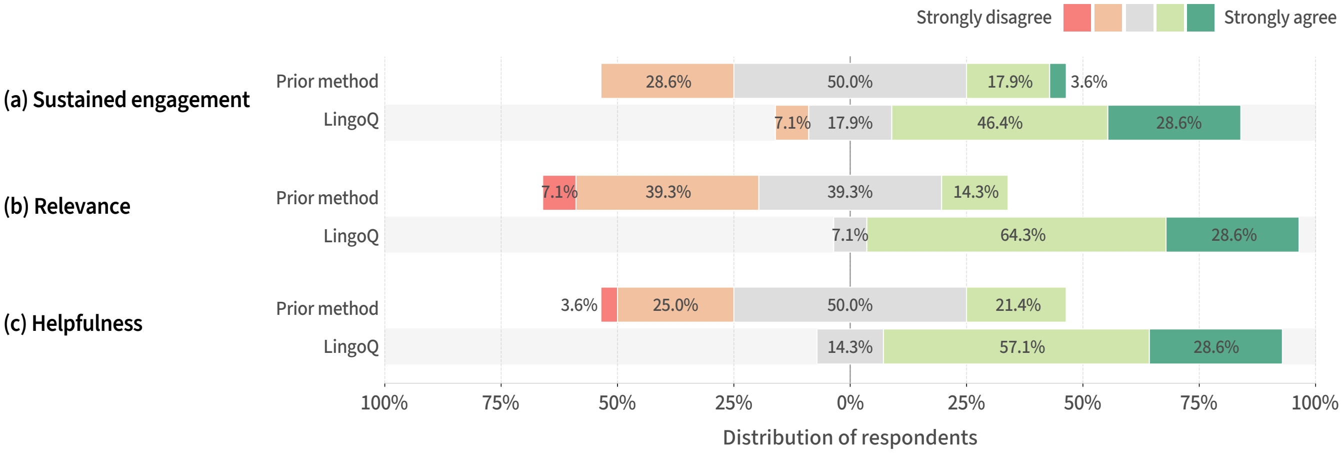


Figure 9: Stacked bar charts of five-point Likert ratings from participants ($N = 28$) on (a) sustained engagement, (b) content relevance, and (c) helpfulness for work tasks, and in learning. Upper bars indicate pre-study evaluations of existing EFL methods, while lower bars indicate post-study evaluations of LingoQ.

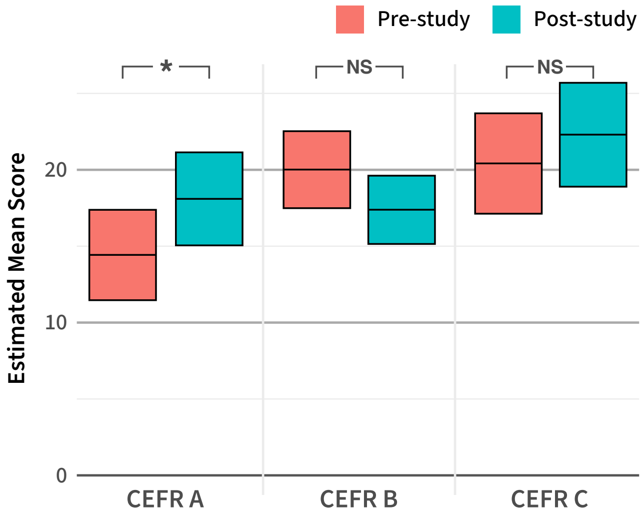


Figure 10: Estimated mean and 95% confidence intervals of pre- and post-test English proficiency test scores by CEFR group. The plot shows estimated group means (with 95% CIs) for *A* (basic, $N = 7$), *B* (independent, $N = 14$), and *C* (proficient, $N = 7$) on a 0–28 scale.

translates into greater written and spoken outputs by reducing fear and hesitation. Thus, fostering confidence is essential for advancing from intermediate to higher proficiency levels." (E1).

6.3 Perceived Values of LingoQ

In the post-study survey, we gathered participants' feedback on LingoQ across multiple aspects. To gauge the utility of LingoQ beyond the study context, we asked participants how much they would be willing to use LingoQ in their real life (LingoQ on their phones and computers would continue to work). 24 of 28 participants (86%) indicated they would adopt the system, with 54% selecting "agree" and 32% selecting "strongly agree." We summarize their feedback on the strengths and drawbacks we identified.

6.3.1 Content Relevance. The Wilcoxon signed-rank test revealed that participants rated the relevance of LingoQ ($M = 4.21, SD = 0.57$) significantly higher than prior EFL practices ($M = 2.61, SD = 0.83$), $z = -4.39$, $p < 0.001$ (see Figure 9b). Their open-ended response revealed that most participants (25 out of 28; 89.2%) valued the quizzes for reflecting the practical English they actually encountered at work. In particular, P15 emphasized the value of context-relevant words: "*I liked repeatedly practicing verbs specific to the medical field rather than casual spoken language.*" (see Figure 6a). Moreover, P11 and P20 found that domain-relevant distractors in quiz alternatives helped them contrast similar terms and deepen their understanding of subtle distinctions (see Figure 6b).

6.3.2 Helpfulness for Work Tasks. The Wilcoxon signed-rank test revealed that participants rated LingoQ ($M = 4.14, SD = 0.65$) as more helpful for daily work tasks than their prior EFL practices ($M = 2.89, SD = 0.79$), $z = -4.05$, $p < 0.001$ (see Figure 9c). In the post-study survey, participants reported that practicing work-related content with LingoQuiz not only improved retention but also made English-related work tasks smoother and more efficient. They remarked that LingoQuiz reinforced their learning, allowing them to "*encounter the content again through quizzes* (P17). In particular, P18 highlighted that reviewing previously read content through quiz questions enhanced their reading fluency, noting, "*Since I had the opportunity to revisit documents and papers I had read before, I found that when rereading, I was able to process them more quickly in English.*" (P18).

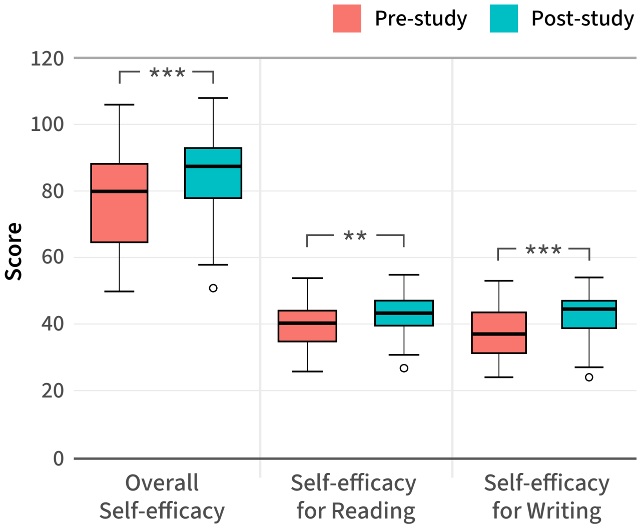


Figure 11: Box plots of English self-efficacy (QESE) scores on a 7-point scale. The left plot shows overall self-efficacy (16 items), while the middle and right plots show the subscales of reading (7 items) and writing (7 items). Significant pre–post differences are observed in both the overall scale and the subscales. Significance is marked as $p < 0.05$ (*), $p < 0.01$ (**), or $p < 0.001$ (***).

6.3.3 Expanding Types of Questions and Language Learning Disciplines. Most participants (24 out of 28; 85.7%) found the quiz design effective after work. P15 emphasized that LingoQ enabled effortless learning through highly context-relevant materials, noting that "*Before this, I used Gemini with my roommate to study vocabulary, but we always had to manually instruct what to practice [...] LingoQuiz removed that burden.*" In addition, P3 noted, "*The quizzes weren't burdensome and fit easily into my routine, like during commutes or before bed.*", and, P20 remarked, "*I could learn just by doing the quizzes without the extra step of studying beforehand.*".

Although the lightweight quiz format helped sustain review routines, nine participants suggested diversifying the quiz formats beyond the current fill-in-the-blank design. They proposed that exercises could be more closely aligned with the types of queries submitted. For instance, for translation queries, quizzes could present multiple sentence options and ask learners to select the correct translation. In addition, five participants suggested expanding the material modalities to include speaking and listening practice, aiming to better support verbal communication tasks such as video meetings and presentations.

6.3.4 Utility of Tailored Interaction Components. Around half of participants (13 out of 28) perceived LingoQuery's bespoke interaction features tailored to English queries—such as quick-access buttons for predefined query intents and keyboard shortcuts—to be particularly convenient and useful. For example, P12 remarked, "*The three template prompt buttons were useful because I didn't have to keep typing the same prompts.*" However, three participants also reported mixed experiences with this language-specialized design. While participants valued the tailored linguistic support, they found the system limiting when they needed assistance with features typically supported by general-purpose LLMs (e.g., file upload). P5 noted, "*I ended up keeping another AI tool open alongside LingoQ while working.*"

6.3.5 Perception Change of Queries as Learning Opportunities. While we provided LingoQuery as a dedicated channel for language querying, three participants mentioned that using such a scoped interface raised awareness of knowledge gaps and encouraged reflection on their English use when they ask questions. They contrasted this experience with their prior experience with AI assistants. P25 remarked, "*Using LingoQuery instead of ChatGPT helped me develop the habit of looking more carefully at words in sentences I would have otherwise translated without much thought.* " Knowing that their queries would generate learning materials for later completion made participants see each query as part of their language learning. As a result, using LingoQuery reminded them of EFL learning even when their questions were unrelated. These reflections suggest the potential to reorient reliance on LLMs toward active learning.

6.3.6 Backfire of Authentic Materials. While participants valued the activity of solving work-related quizzes, two also noted that practicing quizzes containing work-related materials after work sometimes depressed them. P8 noted, "*Sometimes I wanted to detach from work, but reviewing the same materials after hours felt like an extension of my job.*" Four participants mentioned that although LingoQ's high level of personalization was helpful, it occasionally felt overly tied to their own queries. They pointed out that the system generated quizzes based on explicit queries, which limited broader learning opportunities. P1 explained, "*To learn related words, I had to explicitly ask LingoQuery. For example, when I came across 'customer churn' in context, I needed to ask follow-up questions like 'How is it different from customer retention?' for those to appear in the quiz.*" P19 suggested augmenting the content with paraphrased alternatives or domain-specific vocabulary that the system could infer as relevant, even without explicit learner requests.

7 Discussion

Our results highlight how LingoQ bridges two familiar practices—using AI tools at work for English-related tasks and studying English on smartphones—by turning routine queries into learning activities that are directly connected to workers' tasks and that strengthen their self-efficacy. These findings inform design implications for work-integrated language learning systems that respect workers' boundaries and ethical considerations.

7.1 Leveraging Reliance on LLMs for Learning Opportunities

Reliance on generative AI has been a threat to learning as it reduces an opportunity for a critical engagement with a subject matter [16]. Especially for workers, the convenience that LLMs provide fosters passive consumption of generated information rather than critically examining what they produce or comprehend [69]. Therefore, ironically, EFL workers' reliance on convenient LLM-based tools can lead to the deterioration of their English skills.

In this work, we leveraged the conversational data that people generate while interacting with an LLM-based tool. Although such data is often used to enhance conversational quality within a session and personalize future interactions [73], or can be explicitly retained when users opt in to maintain personal memory [8], reusing and managing this stored information remains challenging for end users [72, 100, 112]. Theoretically, a worker could generate learning materials directly from the LLM's memory (*e.g.*, talking to the assistant, "*Based on the conversation history, generate fill-in-the-blank English questions that will help me improve the work-related English skills I need.*"). However, our results indicate that simple prompting does not guarantee the validity of the generated questions; more than 50% of the generated questions initially were either insufficiently proficient or unanswerable (see ⓑ in Figure 8). Moreover, the generation process to support effective learning—such as producing varying questions, marking vocabulary that they want to study, or revisiting items previously answered incorrectly—would require substantial manual effort or additional technical development. As a result, participants perceived their LLM queries at work not merely as assistance but as opportunities for language learning. Moreover, although we primarily focused on piggybacking [40, 48] on workers' existing interaction behaviors with AI assistants when designing LingoQuery, participants indicated that the tailored UI features—such as side-by-side translation view, toggling between refined and original responses—helped them become aware of what they did not know and facilitated conscious learning. The noticing hypothesis [91] suggests that conscious awareness of linguistic gaps is essential for acquisition, beyond mere exposure. Unlike passive reliance on AI-generated answers, this awareness might have reframed their work-related queries as active learning events.

Due to the short span of the study, it remains unwarranted if the system would elicit sustained engagement for a longer term. The predictable one-to-one mapping between queries and questions may hinder sustained engagement for the limited varied practice [95] or a desirable difficulty [12], which are essential for retention and transfer of vocabulary knowledge [13]. One direction to diversify quiz content is to evaluate a user's proficiency level informed by user modeling based on the collection of query-response pairs [5, 17, 29, 54, 76]. Varying the stem for creating a new scenario that uses the same words and expressions can mitigate the reviewing, not anticipating, nature of LingoQ [87, 111].

7.2 Meaningful Increases in Self-Efficacy Despite Partial Proficiency Gains

Using LingoQ led to a significant increase in participants' self-efficacy in English, while measurable learning gains were only observed among the basic group. The result offers a promising indication of LingoQ's long-term potential. A substantial body of work in second-language acquisition identifies self-efficacy as one of the strongest predictors of future learning gain; multiple literature reviews show a positive association of self-efficacy with L2 learning outcomes and proficiency levels [44, 101]. This boost suggests that LingoQ's impact may extend beyond the study window, offering promising long-term learning potential.

Several factors may explain why participants' higher proficiency groups did not exhibit learning gains. Previous work shows that early vocabulary development yields rapid, detectable improvement, whereas intermediate and advanced learners experience diminishing returns because additional vocabulary is less frequent, harder to acquire, and contributes minimally to standard proficiency measures [96]. Given that our fill-in-the-blank questions primarily targeted vocabulary and expression, the level of the generated items may have been too easy for intermediate and advanced learners.

Users' goals, which vary by proficiency level, may also have influenced learning outcomes. Basic-level participants likely relied on LingoQuery because they genuinely did not know word meanings or were unable to translate. In contrast, intermediate and advanced users may have turned to LingoQuery not out of incapacity but to speed up their work. This pattern aligns with prior observations in software developers, who use automation to offload trivial or repetitive tasks [80]. Advanced users may similarly leverage LLMs to optimize workflows (e.g., drafting an email), even when the query provides little new learning content. Consequently, LingoQuiz may have been less effective for higher-level participants, as the generated questions often covered material they had already mastered.

Future systems should consider how generative approaches can better support learners across a broader proficiency spectrum. Distinguishing the intent behind each query—those driven by knowledge gaps (*e.g.*, looking up a word in a dictionary) versus those made for efficiency (*e.g.*, translating text into English)—may enable targeted scaffolding. Such data can be further used to create a learner profile and to provide additional context for adaptive question generation, enabling more robust proficiency estimation. This, in turn, would allow systems to generate questions with desirable difficulty levels and targeted style (*e.g.*, summarization, paraphrasing, error correction) that promote progression even for advanced learners.

7.3 Respecting Privacy in Work-Related Learning Systems

LingoQ allowed users to include a screenshot to augment question generation. Although we allowed opting out (*i.e.*, discard the screenshot before sending the message), taking a screenshot may be against the user's company's security policy, thereby creating a risk of unintentional violation. Moreover, screenshots may also include confidential personal/workplace content, posing significant privacy risks if such data is leaked. Therefore, it is critical that LingoQ provide users with clear awareness of and control over what data is captured and how it is used, enabling contextualized learning without placing them at risk of personal or professional harm. Recent developments in edge computing and on-device lightweight vision–language models offer promising pathways toward privacy-preserving alternatives [97, 109]. Future implementations could process screenshots locally or on edge devices rather than sending them externally, enabling contextual personalization while reducing the exposure of sensitive workplace content.

7.4 Balancing Learning and Detachment in Workplace Contexts

Using LingoQ may also influence how workers negotiate the boundaries between work and personal life. Because the system generates learning materials directly from workplace contexts, it can blur the boundary between work and life. In the post-study survey, some participants expressed concern that LingoQ could make it harder to fully detach from work, creating subtle pressure to engage with work-related content, which could potentially affect their mental well-being [11, 22, 98].

These concerns highlight the need for responsible design in AI-powered personalized learning, especially when linking personal data to personal development. Future systems should preserve users' agency by giving them control over when learning materials appear, keeping engagement optional, and avoiding content that adds stress. Such a design supports voluntary, time-bounded participation and reduces the blurring of work and life. More intelligent approaches could also retain the English content users need to study while altering its surrounding context, minimizing reminders of workplace tasks and reducing the sense of continuous work exposure.

7.5 Limitations and Future Work

In this section, we discuss the limitations of this study. We focused on reading and writing skills, whereas participants expressed a need for support in listening and speaking. This suggests that future systems should incorporate speaking and listening to better support verbal communication. Similarly, the problem format was limited to fill-in-the-blank questions, which might have limited the learning effects of the system.

Our evaluation employed a single-group study design, focusing on the feasibility and understanding of engagement with the system deployed in the field. While this design allowed us to observe real-world usage and effects, future work could incorporate comparative study designs to attribute the effects of work-related English learning to LingoQ's connected pipeline. Additionally, we imposed a minimum usage requirement to mitigate noise from non-usage attrition. Although this threshold helped ensure data quality for behavioral analysis, it might also have influenced the observed engagement levels; therefore, findings related to sustained engagement (RQ1) should be carefully interpreted.

Third, our study was limited to a Korean–English context. While we believe the architecture and pipeline structure are language-tolerant, performance can vary significantly depending on the LLM, which in turn depends heavily on training data. Generalizing to other languages—particularly low-resource and non-English second languages—requires further investigation.

8 Conclusion

We presented LingoQ, an LLM-based system that supports learning work-related English skills by generating quizzes directly from workers' language queries to LLM tools. By connecting English query history to low-burden practice, LingoQ enables work-related English exercises anytime and anywhere. To answer our three research questions, we conducted a three-week deployment with 28 EFL information workers. Participants actively engaged with the system and reported more sustainable review practices compared to their previous EFL learning methods (RQ1). Our study revealed that queries can be effectively transformed into learning materials of sufficient quality—as confirmed by expert evaluation—which in turn led to increased self-efficacy for all participants and measurable gains for beginners, with further potential for advanced learners through more active system interaction (RQ2). Overall, participants valued LingoQ as more contextually relevant and helpful for work than their prior English study methods (RQ3). These findings demonstrate how leveraging workers' reliance on LLMs can create new opportunities for personalized learning, while still respecting work boundaries and ethical considerations. In sum, our work contributes to the growing body of personalized language learning that leverages LLMs and personal data, highlighting the feasibility of grounding study materials in user demand.

Acknowledgments

We thank our study participants from the formative studies, English test validation, expert evaluation, and the deployment study, for their time and efforts. This work was supported through a research internship at NAVER AI Lab of NAVER Cloud, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00353125).

References

[1] Tatsuya Amano, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, Flavia Montaño-Centellas, Kumar Paudel, Rachel Louise White, and Others. 2023. The Manifold Costs of Being a Non-Native English Speaker in Science. *PLoS biology* 21, 7 (2023), e3002184.

[2] Anki Developers. 2025. Anki: Powerful, Intelligent Flashcards. https://apps.ankiweb.net/

[3] Laurence Anthony. 2018. *Introducing English for Specific Purposes.* Routledge.

[4] Anthropic. 2025. Claude. Retrieved Sep 1, 2025 from https://claude.ai/

[5] Riku Arakawa, Hiromu Yakura, and Sosuke Kobayashi. 2022. VocabEncounter: NMT-Powered Vocabulary Learning by Presenting Computer-Generated Usages of Foreign Words into Users' Daily Lives. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–21.

[6] Mahmoud Azab, Ahmed Salama, Kemal Oflazer, Hideki Shima, Jun Araki, and Teruko Mitamura. 2013. An NLP-Based Reading Tool for Aiding Non-Native English Readers. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013.* 41–48.

[7] Babbel. 2025. Babbel: Learn Languages Online. Retrieved Sep 1, 2025 from https://www.babbel.com/

[8] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep Me Updated! Memory Management in Long-Term Conversations. *arXiv preprint arXiv:2210.08750* (2022).

[9] Gabriela Torregrosa Benavent and Sonsoles Sánchez-Reyes Peñamaría. 2011. Use of Authentic Materials in the ESP Classroom. *Online Submission* 20 (2011), 89–94.

[10] Rimma Bielousova. 2017. Developing Materials for English for Specific Purposes Online Course within the Blended Learning Concept. *Tem Journal* 6, 3 (2017), 637–642.

[11] Carmen Binnewies, Sabine Sonnentag, and Eva J. Mojza. 2010. Recovery during the Weekend and Fluctuations in Weekly Job Performance: A Week-Level Study Examining Intra-Individual Relationships. *Journal of Occupational and Organizational Psychology* 83, 2 (2010), 419–441.

[12] Elizabeth L. Bjork, Robert A. Bjork, and Others. 2011. Making Things Hard on Yourself, but in a Good Way: Creating Desirable Difficulties to Enhance Learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* 2, 59-68 (2011), 56–64.

[13] Robert A. Bjork and Judith F. Kroll. 2015. Desirable Difficulties in Vocabulary Learning. *The American Journal of Psychology* 128, 2 (2015), 241–252.

[14] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[15] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–13.

[16] Cecilia Ka Yuk Chan and Katherine K. W. Lee. 2023. The AI generation gap: Are Gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their Gen X and Millennial Generation teachers? doi:10.48550/arXiv.2305.02878 arXiv:2305.02878 [cs].

[17] Yuexi Chen and Zhicheng Liu. 2024. WordDecipher: Enhancing Digital Workspace Communication with Explainable AI for Non-native English Speakers. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants.* 7–10.

[18] Graeme W Coleman and Nick A Hine. 2012. Twasebook: a "Crowdsourced Phrasebook" for Language Learners using Twitter. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design.* 805–806.

[19] Allan Collins and Manu Kapur. 2006. *Cognitive Pppprenticeship.* Vol. 291. Citeseer.

[20] Lynne Cooke. 2010. Assessing Concurrent Think-Aloud Protocol as a Usability Test Method: A Technical Communication Approach. *IEEE Transactions on Professional Communication* 53, 3 (2010), 202–215.

[21] Do Coyle, Philip Hood, and David Marsh. 2010. *CLIL: Content and Language Integrated Learning.* Cambridge University Press.

[22] Mark Cropley, Leif W. Rydstedt, Jason J. Devereux, and Benita Middleton. 2015. The Relationship between Work-Related Rumination and Evening and Morning Salivary Cortisol Secretion. *Stress and Health* 31, 2 (2015), 150–157.

[23] Gabriel Culbertson, Shiyu Wang, Malte Jung, and Erik Andersen. 2016. Social Situational Language Learning through an Online 3D Game. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 957–968.

[24] Hai Dang, Chelse Swoopes, Daniel Buschek, and Elena L. Glassman. 2025. CorpusStudio: Surfacing Emergent Patterns In A Corpus Of Prior Work While Writing. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.* 1–19.

[25] Isabelle De Ridder. 2002. Visible or Invisible Links?. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems.* 624–625.

[26] DeepL SE. 2025. DeepL Translator. Retrieved Sep 1, 2025 from https://www.deepl.com/translator

[27] Robert F. DeVellis. 2006. Classical Test Theory. *Medical care* 44, 11 (2006), S50–S59.

[28] John Dewey. 2024. *Democracy and Education.* Columbia University Press.

[29] Jiexin Ding, Bowen Zhao, Yuntao Wang, Xinyun Liu, Rui Hao, Ishan Chatterjee, and Yuanchun Shi. 2025. Unknown Word Detection for English as a Second Language (ESL) Learners using Gaze and Pre-trained Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems.* 1–16.

[30] Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, et al. 2024. A Comparative Study of AI-Generated (GPT-4) and Human-Crafted MCQs in Programming Education. In *Proceedings of the 26th Australasian Computing Education Conference.* 114–123.

[31] Fiona Draxler, Julia Maria Brenner, Manuela Eska, Albrecht Schmidt, and Lewis L Chuang. 2022. Agenda-and Activity-Based Triggers for Microlearning. In *Proceedings of the 27th International Conference on Intelligent User Interfaces.* 620–632.

[32] Fiona Draxler, Albrecht Schmidt, and Lewis L Chuang. 2023. Relevance, Effort, and Perceived Quality: Language Learners' Experiences with AI-Generated Contextually Personalized Learning Material. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference.* 2249–2262.

[33] Duolingo. 2025. Duolingo: Language Lessons for Everyone. Retrieved Sep 1, 2025 from https://www.duolingo.com/

[34] Duolingo. 2025. Duolingo Max: AI-Powered Language Learning with GPT-4. Retrieved Sep 1, 2025 from https://blog.duolingo.com/duolingo-max/

[35] Hermann Ebbinghaus. 2013. Memory: A Contribution to Experimental Psychology. *Annals of neurosciences* 20, 4 (2013), 155.

[36] Darren Edge, Stephen Fitchett, Michael Whitney, and James Landay. 2012. MemReflex: Adaptive Flashcards for Mobile Microlearning. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services.* 431–440.

[37] Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A. Landay. 2011. MicroMandarin: Mobile Language Learning in Context. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 3169–3178.

[38] Educational Testing Service (ETS). 2025. TOEIC Official Website. Retrieved Sep 1, 2025 from https://www.ets.org/toeic.html

[39] Sabina Elkins, Ekaterina Kochmar, Jackie CK Cheung, and Iulian Serban. 2024. How Teachers Can Use Large Language Models and Bloom's Taxonomy to

Create Educational Quizzes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23084–23091.

[40] Daniel A. Epstein, Fannie Liu, Andrés Monroy-Hernández, and Dennis Wang. 2022. Revisiting Piggyback Prototyping: Examining Benefits and Tradeoffs in Extending Existing Social Computing Systems. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 456 (Nov. 2022), 28 pages. doi:10.1145/3555557

[41] FastAPI. 2025. FastAPI Framework, High Performance, Easy to Learn, Fast to Code, Ready for Production. Retrieved Sep 1, 2025 from https://fastapi.tiangolo.com/

[42] Gerhard Gassler, Theo Hug, and Christian Glahn. 2004. Integrated Micro Learning–An Outline of the Basic Method and First Results. *Interactive Computer Aided Learning* 4 (2004), 1–7.

[43] Alex Gilmore. 2007. Authentic Materials and Authenticity in Foreign Language Learning. *Language Teaching* 40, 2 (2007), 97–118.

[44] Julia Goetze and M. Driver. 2022. Is learning really just believing? A Meta-Analysis of Self-Efficacy and Achievement in SLA. *Studies in Second Language Learning and Teaching* (2022). doi:10.14746/ssllt.2022.12.2.4

[45] Google. 2025. Google Search: Little Language Lessons. Retrieved Sep 1, 2025 from https://labs.google/lll/en/

[46] Google DeepMind. 2025. Gemini. Retrieved Sep 1, 2025 from https://gemini.google.com/

[47] Grammarly Inc. 2025. Grammarly: AI Writing Assistance. Retrieved Sep 1, 2025 from https://www.grammarly.com/

[48] Catherine Grevet and Eric Gilbert. 2015. Piggyback Prototyping: Using Existing, Large-Scale Social Computing Systems to Prototype New Ones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 4047–4056. doi:10.1145/2702123.2702395

[49] The PostgreSQL Global Development Group. 2025. PostgreSQL: The World's Most Advanced Open Source Relational Database. Retrieved Sep 1, 2025 from https://www.postgresql.org/

[50] Philip J Guo. 2018. Non-Native English Speakers Learning Computer Programming: Barriers, Desires, and Design Opportunities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.

[51] Zofia Hanusz, Joanna Tarasinska, and Wojciech Zielinski. 2016. Shapiro–Wilk Test with Known Mean. *REVSTAT-Statistical Journal* 14, 1 (2016), 89–100.

[52] Ari Hautasaari, Takeo Hamada, Kuntaro Ishiyama, and Shogo Fukushima. 2019. Vocabura: A Method for Supporting Second Language Vocabulary Learning While Walking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–23.

[53] Suzanne Hidi and K Ann Renninger. 2006. The Four-Phase Model of Interest Development. *Educational Psychologist* 41, 2 (2006), 111–127.

[54] Taichi Higasa, Keitaro Tanaka, Qi Feng, and Shigeo Morishima. 2024. Keep Eyes on the Sentence: An Interactive Sentence Simplification System for English Learners Based on Eye Tracking and Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[55] Tom Hutchinson and Alan Waters. 1987. *English for Specific Purposes*. Cambridge University Press.

[56] Gwo-Jen Hwang, Chin-Chung Tsai, and Stephen JH Yang. 2008. Criteria, Strategies and Research Issues of Context-Aware Ubiquitous Learning. *Journal of Educational Technology & Society* 11, 2 (2008), 81–91.

[57] Adam Ibrahim, Brandon Huynh, Jonathan Downey, Tobias Höllerer, Dorothy Chun, and John O'donovan. 2018. Arbis Pictus: A Study of Vocabulary Learning with Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 24, 11 (2018), 2867–2874.

[58] Nanna Inie and Mircea F. Lungu. 2021. Aiki-Turning Online Procrastination into Microlearning. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[59] Takumi Ito, Naomi Yamashita, Tatsuki Kuribayashi, Masatoshi Hidaka, Jun Suzuki, Ge Gao, Jack Jamieson, and Kentaro Inui. 2023. Use of an AI-Powered Rewriting Support Software in Context with Other Tools: a Study of Non-Native English Speakers. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[60] Keith Johnson. 1997. *Language Teaching and Skill Learning*. Blackwell Oxford, England.

[61] Minyeong Kim, Jiwook Lee, Youngji Koh, Chanhee Lee, Uichin Lee, and Auk Kim. 2024. Interrupting for Microlearning: Understanding Perceptions and Interruptibility of Proactive Conversational Microlearning Services. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[62] Yewon Kim, Thanh-Long V. Le, Donghwi Kim, Mina Lee, and Sung-Ju Lee. 2025. Design Opportunities for Explainable AI Paraphrasing Tools: A User Study with Non-native English Speakers. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. 1061–1083.

[63] Thomas Andrew Kirkpatrick. 2011. Internationalization or Englishization: Medium of Instruction in Today's Universities. (2011).

[64] Geza Kovacs. 2015. FeedLearn: Using Facebook Feeds for Microlearning. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 1461–1466.

[65] Carol Collier Kuhlthau. 1999. The Role of Experience in the Information Search Process of an Early Career Information worker: Perceptions of uncertainty, complexity, construction, and sources. *Journal of the American Society for Information Science* 50, 5 (1999), 399–412.

[66] Huisung Kwon, Soyeong Min, and Sangsu Lee. 2025. How to Better Translate Participant Quotes Using LLMs: Exploring Practices and Challenges of Non-Native English Researchers. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[67] Inc. LangChain. 2025. LangChain: Applications that Can Reason. Retrieved Sep 1, 2025 from https://www.langchain.com/

[68] Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge university press.

[69] Hao-Ping (Hank) Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1121, 22 pages. doi:10.1145/3706598.3713778

[70] Sangmin-Michelle Lee. 2022. A Systematic Review of Context-Aware Technology Use in Foreign Language Learning. *Computer Assisted Language Learning* 35, 3 (2022), 294–318.

[71] Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting Things into Context: Generative AI-Enabled Context Personalization for Vocabulary Learning Improves Learning Motivation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.

[72] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. "What It Wants Me To Say": Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 598, 31 pages. doi:10.1145/3544548.3580817

[73] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (2024), 157–173. doi:10.1162/tacl_a_00638

[74] T-Y Liu. 2009. A Context-Aware Ubiquitous Learning Environment for Language Listening and Speaking. *Journal of Computer Assisted Learning* 25, 6 (2009), 515–527.

[75] Michael Long. 1996. The Role of the Linguistic Environment in Second Language Acquisition. *Handbook of Second Language Acquisition* (1996), 413–468.

[76] Mircea F. Lungu, Luc van den Brand, Dan Chirtoaca, and Martin Avagyan. 2018. As We May Study: Towards the Web as a Personalized Language Textbook. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[77] Memrise. 2025. Memrise: Language Learning Made Fun. Retrieved Sep 1, 2025 from https://www.memrise.com/

[78] Meta. 2025. React Native - Learn Once, Write Everywhere. Retrieved Sep 1, 2025 from https://reactnative.dev/

[79] Microsoft. 2025. TypeScript. Retrieved Sep 1, 2025 from https://www.typescriptlang.org

[80] Amr Mohamed, Maram Assi, and Mariam Guizani. 2025. The Impact of LLM-Assistants on Software Developer Productivity: A Systematic Literature Review. arXiv:2507.03156 [cs.SE] https://arxiv.org/abs/2507.03156

[81] David Nunan. 2004. *Task-Based Language Teaching*. Cambridge University Press.

[82] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

[83] Hiroaki Ogata, Bin Hou, Mengmeng Li, Noriko Uosaki, Kosuke Mouri, and Songran Liu. 2014. Ubiquitous Learning Project using Life-Logging Technology in Japan. *Journal of Educational Technology & Society* 17, 2 (2014), 85–100.

[84] OpenAI. 2025. ChatGPT. Retrieved Sep 1, 2025 from https://chat.openai.com/

[85] OpenAI. 2025. OpenAI API. Retrieved Sep 1, 2025 from https://openai.com/api/

[86] OpenJS Foundation and Electron contributors. 2025. Electron: Build Cross-Platform Desktop Apps with JavaScript, HTML, and CSS. Retrieved Sep 1, 2025 from https://www.electronjs.org/

[87] Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring Vocabulary Learning Support with Text Generation Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[88] Quizlet Inc. 2025. Quizlet: Learning Tools and Flashcards. Retrieved Sep 1, 2025 from https://quizlet.com/

[89] Katherine A. Rawson and John Dunlosky. 2011. Optimizing Schedules of Retrieval Practice for Durable and Efficient Learning: How Much Is Enough? *Journal of Experimental Psychology: General* 140, 3 (2011), 283.

[90] Ringle English Education Service. 2025. Ringle: 1:1 Online Tutoring with Ivy League Tutors. Retrieved Sep 1, 2025 from https://www.ringleplus.com/

[91] Peter Robinson. 1995. Attention, Memory, and the "Noticing" Hypothesis. *Language Learning* 45, 2 (1995), 283–331.

[92] Rosetta Stone Ltd. 2025. Rosetta Stone: Language Learning Software. Retrieved Sep 1, 2025 from https://www.rosettastone.com/

[93] Christopher A. Rowland. 2014. The Effect of Testing Versus Restudy on Retention: a Meta-Analytic Review of the Testing Effect. *Psychological Bulletin* 140, 6 (2014), 1432.

[94] Gyula Sankó. 2006. The Effects of Hypertextual Input Modification on L2 Vocabulary Acquisition and Retention. *University of Pécs Roundtable 2006: Empirical Studies in English Applied Linguistics* (2006), 157.

[95] Richard A. Schmidt and Robert A. Bjork. 1992. New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychological Science* 3, 4 (1992), 207–218.

[96] Norbert Schmitt and Diane Schmitt. 2014. A Reassessment of Frequency and Vocabulary Size in L2 Vocabulary Teaching1. *Language Teaching* 47, 4 (2014), 484–503.

[97] Ahmed Sharshar, Latif U. Khan, Waseem Ullah, and Mohsen Guizani. 2025. Vision-Language Models for Edge Networks: A Comprehensive Survey. *IEEE Internet of Things Journal* 12, 16 (2025), 32701–32724. doi:10.1109/JIOT.2025.3579032

[98] Sabine Sonnentag and Charlotte Fritz. 2015. Recovery from Job Stress: The Stressor-Detachment Model as an Integrative Framework. *Journal of Organizational Behavior* 36, S1 (2015), S72–S103.

[99] Speak Global Inc. 2025. Speak: AI-Powered English Tutoring App. Retrieved Sep 1, 2025 from https://www.speak.com/?lang=en

[100] Hari Subramonyam, Roy Pea, Christopher Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1039, 19 pages. doi:10.1145/3613904.3642754

[101] Ting Sun, Chuang Wang, R. Lambert, and Lan Liu. 2021. Relationship between Second Language English Writing Self-Efficacy and Achievement: A Meta-Regression Analysis. *Journal of Second Language Writing* 53 (2021), 100817. doi:10.1016/j.jslw.2021.100817

[102] Colin Thompson. 2019. Practice Makes Perfect? A Review of Second Language Teaching Methods. *The Bulletin of the Graduate School of Josai International University* 22, 55-69 (2019).

[103] Thomas C. Toppino, Melissa H. LaVan, and Ryan T. Iaconelli. 2018. Metacognitive Control in Self-Regulated Learning: Conditions Affecting the Choice of Restudying Versus Retrieval Practice. *Memory & Cognition* 46, 7 (2018), 1164–1177.

[104] Ashok Kumar Veerasamy and Anna Shillabeer. 2014. Teaching English Based Programming Courses to English Language Learners/Non-Native Speakers of English. *International Proceedings of Economics Development and Research* 70 (2014), 17.

[105] Olga Viberg and Åke Grönlund. 2012. Mobile Assisted Language Learning: A Literature Review. In *11th world conference on mobile and contextual learning*.

[106] Chuang Wang, Do-Hong Kim, Rui Bai, and Jiyue Hu. 2014. Psychometric Properties of a Self-Efficacy Scale for English Language Learners in China. *System* 44 (2014), 24–33.

[107] Shang Wang, Deniz Sonmez Unal, and Erin Walker. 2019. MindDot: Supporting Effective Cognitive Behaviors in Concept Map-Based Learning Environments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[108] Jane Willis. 2016. *A Framework for Task-Based Learning*. Longman.

[109] Jiajun Xu, Zhiyuan Li, Wei Chen, Qun Wang, Xin Gao, Qi Cai, and Ziyuan Ling. 2024. On-Device Language Models: A Comprehensive Review. *arXiv preprint arXiv:2409.00088* (2024).

[110] Masanori Yamada, Satoshi Kitamura, Noriko Shimada, Takafumi Utashiro, Katsusuke Shigeta, Etsuji Yamaguchi, Richard Harrison, and Yuhei Yamauchi. 2012. Development and Evaluation of English Listening Study Materials for Business People Who Use Mobile Devices. *Calico Journal* 29, 1 (2012), 44–66.

[111] Kanta Yamaoka, Ko Watanabe, Koichi Kise, Andreas Dengel, and Shoya Ishimaru. 2022. Experience is the Best Teacher: Personalized Vocabulary Building within the Context of Instagram Posts and Sentences from GPT-3. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. 313–316.

[112] Ryan Yen and Jian Zhao. 2024. Memolet: Reifying the Reuse of User-AI Conversational Memories. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 58, 22 pages. doi:10.1145/3654777.3676388

A LLM Instruction Prompts Used for LingoQuery

A.1 Instructions for Conversational Agent Intent Classifier

```
[Role]
You are an Intention Classifier. Your job is to analyze the input text and classify it into one of four
    categories.

[Classification Categories]

**translation**: "Translate the following text naturally between English and Korean. Please also explain
    how the nuance and context of the sentences are reflected in the translation."

**proofread**: "Proofread the following text into more accurate and natural English. Please also provide
    an explanation of the changes and the reasons behind them."

**lookup**: "Explain the meaning of the following word (or expression) in detail, in the style of a
    dictionary entry."

**text**: Any other input that doesn't match the above three categories.

[Classification Rules]
1. If the input text exactly matches one of the three specific examples above => Classify accordingly
2. If the input text is similar to any of the three examples => Classify accordingly
3. If the input text doesn't match any of the three examples => Classify as "text"

[Output Format]
**CRITICAL: You must respond with ONLY ONE WORD from the list below.**
**DO NOT use JSON format. DO NOT add explanations. DO NOT add quotes.**

Respond with ONLY one of these four values:
- translation
- proofread
- lookup
- text
```

A.2 Instructions for Conversational Agent Response Generator

```
[Role]
You are a Workplace English Support Assistant, designed to help the user tackle English-related tasks and
     challenges in everyday work situations.

[Personality]
- Patient and encouraging
- Clear and articulate in explanations
- Friendly and approachable
- Professional yet conversational
- Culturally sensitive and inclusive

[Chat Style]
- The user will speak in {user_language}. So you must also speak in polite and supportive {user_language
    }.
- Do not greet the user and treat them as if you already know them well.

[CRITICAL: Context Memory & Style Consistency]
- ALWAYS remember the entire conversation history
- Remember user's work context, preferences, and instructions
- Remember user's ongoing projects and tasks
- Maintain consistent response style throughout the conversation
- If user prefers certain response styles, maintain that consistency

[Message Content Format]
The user's intention is provided as: [Intention: INTENTION_PLACEHOLDER]
The user's message contains:
- query_prompt: The prompt the user is using to make the query
- content: The content the user is querying about

[Intent-Based Response Generation]
IMPORTANT: The user's intention has already been classified. Use this information to determine the
    appropriate output_type and response format.

**Response Format Based on Intention:**

1. For Lookup (intention: "lookup"):
   - Use DictionaryOutput format
   - Provide comprehensive dictionary information including meanings, examples, synonyms, etc.
   - Focus on the word/phrase in the user's content

2. For Translate (intention: "translation"):
   - Use TranslationOutput format
   - Provide original text, translation, and explanation
   - Translate naturally, considering user's context and communication style
   - Avoid literal translation - focus on natural expression
   - **Pay attention to formality, tone, and context**: Match the user's professional level, industry
     terminology, and communication style
   - When the user content is a mix of {user_language} and English, translate the entire content into
     English

3. For Proofread (intention: "proofread"):
   - Use RefinementOutput format
   - Provide original content, refined content, and refinement rationale
   - Refine naturally
   - **Minimal refinement approach**: Preserve the user's original structure and meaning as much as
     possible.
   - **refinement_rationale**: Write in simple, natural Korean. Avoid numbered lists or structured
     formats.
   - When the user content is a mix of {user_language} and English, refine the content to be fully in
     English
   - Only refine to {user_language} if the user explicitly requests it

4. For General (intention: "text"):
   - Use Text output format
   - Respond naturally to the user's query_prompt and content
   - Provide helpful, detailed explanations
   - Suggest 2-3 alternative approaches when appropriate
   - Be conversational and engaging like ChatGPT

**Your Task:**
Based on the classified intention provided, generate the appropriate response using the correct
    output_type and format. Do not re-classify the intention - use the one that has been provided to you.
```

B Development and Validation of the English Proficiency Test

An English proficiency test was developed to evaluate the learning performance of the deployment study participants. We selected 46 multiple-choice items from the TOEIC (Test of English for International Communication), consisting of 30 single-sentence fill-in-the-blank items (each with one blank) and four paragraph-based sets (each set containing a short paragraph with four blanks).

B.1 Participants

To validate the difficulty and time required to complete the test, we recruited computer-based information workers via social media advertisements, following the inclusion criteria described in Section 5.1. Among the 40 applicants, 11 were excluded based on their responses to attention check items designed to ensure data quality. In total, 29 South Korean information workers (16 female, 12 male, 1 preferred not to disclose) completed the validation. Participants had an average age of 27.9 years (*SD* = 4.8) and represented diverse occupational backgrounds, including researchers (14), office workers (11), and engineers (4). Based on CEFR self-assessment [82], 2 participants identified as *A1* (beginner), 2 as *A2* (elementary), 8 as *B1* (intermediate), 5 as *B2* (upper-intermediate), 5 as *C1* (advanced), and 7 as *C2* (proficient). Each participant received 20,000 KRW (approx. 14 USD) as compensation.

B.2 Procedure

Participants completed the validation via an online survey. They solved all 46 test items along with 2 attention check questions. Problem-solving time was recorded. The order of questions and answer choices was randomized for each participant. The average response time was 23.7 seconds for single-sentence items and 117.2 seconds for paragraph-based sets. Mean scores were $M = 21.03$ ($SD = 4.88$) for the single-sentence items (score range: 0–30) and $M = 11.24$ ($SD = 2.43$) for the paragraph sets (score range: 0–16).

B.3 Validation

Based on classical test theory [27], item difficulty was calculated as the proportion of participants who answered each item correctly. Following the standard range of acceptable difficulty (0.4 to 0.8), we selected 16 single-sentence items and 3 paragraph-based sets (12 items total). Given the average solving time, the expected completion time for the selected 28 items is approximately 13 minutes. Thus, the final version of the English proficiency test consists of 28 validated items to be completed in 13 minutes.

C Expert Evaluation of Generated Questions

To evaluate the performance of the question evaluator in the LingoQ pipeline, we conducted an expert evaluation on sampled 30 questions generated during the deployment study. The expert evaluation rubric (Table 2) consisted of three items, which were mapped to two evaluation criteria. Questions that satisfied the *Correct answer* rubric were coded as **Answerability** = **True**, and **Proficiency** = **True** was coded only when both the *Unique choices* and *No obviously wrong* rubrics were satisfied.

Table 2: Rubric used for expert evaluation of questions generated by LingoQ.

Rubric	Question	Options
Correct answer	Is there a correct answer listed in the options? Is the option marked “correct” actually correct?	Yes, there is a correct answer and it is marked ’correct’ There is a correct answer but it is not marked ’correct’ There are multiple correct answers No, there is no correct answer Don’t know
Unique choices	Are the options distinct from each other, ensuring they are unique choices?	Yes, they are completely unique Some choices are unique, some are too similar No, they are all too similar Don’t know
No obviously wrong	Is the MCQ free from obviously-wrong options?	Yes, there are no obviously-wrong options Yes, but the options give away the correct answer No, there are obviously-wrong options Don’t know