Good Fences Make Good Learning: How Self-Directed Language Learners Navigate LLM Delegation Decisions

Jiwon Song
jwsong@hcil.snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Aeri Cho
archo@hcil.snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Sihyeon Lee
sihyeon@hcil.snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Kiroong Choe
krchoe@hcil.snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Jinwook Seo
jseo@hcil.snu.ac.kr
Seoul National University
Seoul, Republic of Korea

Abstract

Self-directed language learners increasingly turn to large language models (LLMs) for assistance, but face the challenge of deciding what learning tasks to delegate to LLMs and how. While prior research has examined the effectiveness of LLM in improving language proficiency, less is known about how learners negotiate agency and what values guide delegation strategies. To address this gap, we conducted a two-part study: an analysis of discussions in the r/languagelearning subreddit to map learners' LLM usage patterns and factors driving delegation, followed by a technology probe study where learners designed learning activities and experimented with LLM support. Our findings reveal three key considerations influencing delegation: accuracy, independence, and authenticity. We analyze these considerations through two types of obstacles: selection challenges in choosing appropriate strategies and execution challenges in following through on intentions. These insights inform the design of AI-assisted learning systems that preserve learner agency while supporting diverse learning goals.

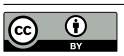
CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in collaborative and social computing*.

Keywords

Learning, Self-directed learning, Language learning, LLM, Large Language Model, AI Collaboration

ACM Reference Format:
Jiwon Song, Aeri Cho, Sihyeon Lee, Kiroong Choe, and Jinwook Seo. 2026. Good Fences Make Good Learning: How Self-Directed Language Learners Navigate LLM Delegation Decisions. 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, 18 pages. https://doi.org/10.1145/3772318.3791657



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791657

1 Introduction

Self-directed learning refers to a learning approach in which learners design their own learning journey, encompassing planning, execution, monitoring, and evaluation. Compared with traditional instructor-centered education, this approach offers learners greater freedom and flexibility to pursue their individual preferences, but it also places greater responsibility and risk on them for learning outcomes. This responsibility intensifies in foreign language learning, one of the most popular yet most demanding domains for self-directed learning. Language proficiency spans multiple skills: vocabulary and grammar knowledge, reading and listening comprehension, speaking and writing production. Self-directed language learners must manage multiple learning threads that require different resources and strategies, which they often lack without institutional guidance.

The recent rise of large language models (LLMs) has opened new opportunities for self-directed language learners by providing unprecedented access to personalized explanations, instant feedback, and practice opportunities. Yet successful integration depends on how learners construct delegation boundaries: what tasks they delegate to LLMs and what they retain for themselves. However, constructing effective boundaries poses significant challenges. Building on prior work examining discrepancies between strategic knowledge and action [25], which distinguishes mediation deficiency (lacking knowledge to identify beneficial strategies) from production deficiency (knowing what's beneficial but failing to apply it), we identify analogous challenges in LLM-assisted learning. *Selection challenge* arises when learners unreflectively or mistakenly choose suboptimal delegation strategies that undermine their learning goals. *Execution challenge* occurs when learners know what they should do but struggle to implement their intentions.

Language learning systems aim to help address these challenges, but effective designs require understanding how learners reason about delegation—which activities they view as essential to retain, which they willingly offload to LLMs, and what values guide these distinctions. Designing learning assistance without these foundations may impose assumptions that override learner autonomy and fail to support the diverse goals that motivate individual learners. Yet existing approaches have focused primarily on technological novelty (e.g., LLM capabilities for language learning assistance) [61] or on validating educational values in teacher-mediated settings.

How self-directed language learners construct and maintain delegation boundaries remains underexplored. To provide this foundation, we conduct a two-part study that addresses the following research questions:

- **RQ1**: How do self-directed language learners construct delegation boundaries in LLM-assisted learning?
- **RQ2**: What underlying values and learning contexts shape these delegation strategies?

To explore the decision space and major rationales for LLM-assisted language learning, we analyzed online community discussions on the r/languagelearning subreddit of Reddit, one of the most active communities for language learners. Our analysis revealed five major tasks where language learners leverage LLMs: planning, conceptual explanations, language input practice, language output practice, and evaluation. Learners expressed diverse opinions on the appropriate use of LLMs, and these decisions were largely shaped by three central considerations: ACCURACY, INDEPENDENCE, and AUTHENTICITY. ACCURACY issues are especially critical in the learning context, where users often lack the ability to reliably discern hallucinations. INDEPENDENCE concerns arise because LLM delegations may undermine learning by replacing the deliberate effort required for skill development. Finally, AUTHENTICITY emerged as a consideration unique to language learning, as learners view human communication as a central element in language use.

Building upon these findings, we conducted a technology probe study to collect and analyze individual processes of LLM involvement that could not be fully captured through online discussions. Our agent design and customization were informed by the delegated tasks and the decision space of Reddit users. The interview questions were designed based on the tensions surrounding the three considerations. After completing three learning sessions with LLM support provided by the probe system, they were interviewed on their overall LLM usage strategies and their desired forms of assistance. From the study results, we derived insights on how the three key considerations interact with individual contexts to shape delegation decisions. While the participants were well aware of the hallucination risks, their reactions were greatly influenced by individual learning priorities. Participants were generally confident in their ability to maintain INDEPENDENCE, but were troubled by the additional burden of learning design and control. Participants showed mixed opinions and reactions towards the AUTHENTICITY of LLM interactions. We interpret these findings under the framework of execution challenge and selection challenge, then provide design implications to help handle these difficulties.

This study presents the following contributions to the field of HCI:

- We reveal how self-directed language learners construct AI delegation strategies by balancing multiple considerations such as ACCURACY, INDEPENDENCE, and AUTHENTICITY within their individual contexts.
- We propose design implications for AI-assisted language learning systems that help self-directed learners make delegation decisions aligned with their individual learning contexts and personal goals.

2 Related Work

2.1 Self-Directed Learning in Languages

Self-directed learning (SDL) emerged to capture the uniqueness of the learner-centered learning experience, as opposed to the traditional research focus of pedagogy and instructor-centered education [8]. Knowles defines SDL in the broadest sense as "a process in which individuals take the initiative, with or without the help of others, in diagnosing their learning needs, formulating learning goals, identifying resources for learning, choosing and implementing appropriate learning strategies, and evaluating learning outcomes [39]". SDL differs with self-regulated learning (SRL) in that SDL highlights learner autonomy over the whole learning journey (setting objectives, selecting resources, evaluating outcomes), whereas SRL emphasizes ongoing regulation of strategies and effort within a given task or curriculum [52, 55]. While SDL may involve SRL processes, SDL provides a more comprehensive model for understanding learners who take agency over their own learning.

While SDL does not necessarily exclude instructor or institution guidance unlike in informal learning, learners assume primary responsibility as decision-makers throughout their learning process [91]. This creates additional metacognitive burdens for self-directed learners, as they assume dual roles: learners of the target content and architects of their learning process [20, 65]. These metacognitive demands intensify in language learning contexts as the domain's complexity requires learners to coordinate multiple skill areas of reading, listening, speaking, and writing simultaneously [31, 85]. These skills demand distinct approaches, resources, and strategies, introducing a new decision space for learners to navigate. Learners should evaluate their current status, decide on priorities, and adopt appropriate resources or strategies for different skill areas based on their individual goals, strengths, and progress [73].

The complexity of these decisions has intensified in modern digital learning environments. Initially, the major obstacle in SDL was resource scarcity compared to traditional teacher-led classrooms. However, the digital revolution has transformed this landscape from one of resource scarcity to one of choice overload. Modern self-directed language learners now face an array of learning platforms, tools, and resources, each offering different approaches to language acquisition [15]. According to Candy, SDL is not just a process of learning the subject matter, but also a goal [13]: learners should cultivate metacognitive learning skills and self-determination [58]. Ironically, this proliferation of tools may undermine the development of learner autonomy when learners become overwhelmed by choice or dependent on external guidance. This tension intensifies when powerful AI tools enter the learning ecosystem, fundamentally transforming the nature of decisions learners must navigate. Our work aims to provide foundation for designs that not only target language learning efficiency but also aims to assist users in taking agency of the learning process. This calls for an alternative approach that focuses on the decision making process within a more open-ended system without strong interventions.

2.2 LLM integration in Self-Directed Language Learning

The term computer-assisted language learning (CALL) first emerged in 1983, and went through three phases: behavioristic, cognitive, and integrative [83]. Beginning as simple drill-and-practice programs for habit formation (behavioristic phase), CALL has evolved with advances in computer technology and educational theories, moving to interaction-focused conversation exercises (cognitive phase) and then to a more socio-cognitive view (integrative phase). The introduction of artificial intelligence technology led to new attempts, known as intelligent tutoring systems (ITS), which aim to approximate the benefits of individualized human tutoring in a scalable, computer-based form [56]. ITS draws from learning science and cognitive theories to provide a pedagogically effective learning process. A good example is the cognitive tutor, designed based on Anderson's Adaptive Control of Thought-Rational (ACT-R) theory [4].

After the arrival of LLMs, the field of language learning has been enthusiastic to explore the potential of the new technology. While previous approaches were pedagogically reliable, the highly structured systems were limited in terms of individual adaptability and natural language processing performance. Integrated AIs were often task-specific with limited roles in the learning process [14]. Student-to-AI interaction was uncommon in AI-based language learning systems as they focused on teacher-AI interactions or unidirectional output from AI to students [14]. The introduction of versatile LLMs fundamentally transformed this landscape, shifting decisions from "which specialized tool for which task?" to "what should I delegate to this multi-purpose AI?" LLM's ability to adapt and personalize interactions improves language learning outcomes by sustained engagement and tailored assistance [2].

LLMs serve multiple roles in language learning [42, 89]: as tutors guiding writing iterations [7], supporting formal language learning [67]. Other common directions involved LLMs as feedback providers evaluating student outputs and suggesting improvements [6, 29] and as generators of customized learning resources [48, 49], including short stories [34] and technical terms [1]. LLMs also provided higher-level support, such as setting personalized goals and detailed lesson plans [30, 44, 49].

However, language learning with LLM assistance entails new risks and concerns. The major concerns include output quality issues, such as inaccurate information [1, 26, 51], and bias toward standard language [62], which require learners to develop fact-checking and evaluation skills [35]. Li et al. identify structural challenges, including a "learning optimization gap" where learners struggle to leverage LLM affordances and a "knowledge comprehension gap" between AI-generated content and learners' integration capacity [43]. Cognitive offloading remains a persistent challenge [42] as students often focus on LLM outputs rather than skill development [1, 54, 86]. There was also evidence of LLMs frequently falling short in motivation[1, 29], engagement [34, 54, 86], and practice opportunities [34, 87]. These challenges develop into a complex decision-making problem: when using AI assistance, writers face a dilemma between imitation for learning and plagiarism [84]. Weaker writers often failed to maintain balance and overrelied on AI, hindering the development of foundational skills.

The main challenge of self-directed language learners in using LLMs, therefore, is not on boolean adoption of LLMs, but on determining the optimal delegation boundaries that preserve learning effectiveness while leveraging AI capabilities [28]. This navigation involves developing meta-AI skills like prompt engineering and output evaluation [68] and working within institutional guidance frameworks [49]. In the face of these challenges, self-directed learners must balance immediate AI benefits against the SDL goal of developing autonomous learning capabilities [32], without institutional scaffolding to guide their choices. While AI systems targeting these learners aim to assist LLM management and learning optimization, their design requires an understanding of the natural preferences and decision-making of their users.

However, current research remains limited in its examination of learner-side decision-making regarding LLM utilization in self-directed language learning, especially how learners develop and justify task-wise delegation boundaries. B. Li et al. interviewed YouTubers who own language learning channels to examine how ChatGPT redefines self-directed learning, extending Song and Hill's SDL framework [70] to include both local factors (personal traits and adaptive learning processes) and global factors (evolving AI technology and sociocultural contexts) [44] that influence college students' SDL. Z. Li et al. build upon Garrison's model of SDL [27] to investigate the general motivation, self-management strategies, and self-monitoring strategies of self-directed language learners in ChatGPT use [47].

Although acknowledging task-related perceptions, these studies do not explore how learners actively construct delegation boundaries through contextual reasoning about individual situations and values. Designs without knowledge of natural usage boundaries may deviate from actual needs, compromise learner agency, and, as a result, undermine individual goals. In this study, we aim to provide an empirical analysis of the delegation decision-making of self-directed language learners to inform future designs of LLM-based language learning assistance.

2.3 LLM Delegation Strategies and Frameworks

The versatility of LLMs has attracted much research interest for delegation frameworks in general domains. Unlike task-specific tools with fixed functions, LLMs enable users to dynamically adapt delegation patterns based on context [60, 81]. Lubars and Tan [53] proposed a foundational framework for understanding task delegability, identifying four key factors: motivation, difficulty, risk, and trust, with trust emerging as the most critical factor and users preferring human-in-the-loop designs over full automation.

Recent studies have shown how people manage this complexity through selective delegation principles based on their mental models. Users readily delegate information-seeking tasks while maintaining control over complex analytical work [37, 53, 69, 77]. Buçinca et al. [12] demonstrated that people develop general heuristics about delegation decisions, though cognitive forcing functions are needed to prevent overreliance on inaccurate outputs. Lai et al. [40] introduced conditional delegation, where humans create explicit rules to define trustworthy regions of AI models, enabling scalable yet selective automation for high-volume tasks.

However, there exist deeper tensions that complicate AI use strategies. Tankelevitch et al. [72] found that generative AI systems impose significant metacognitive demands, requiring users to balance the benefits of automation with the cognitive effort required for effective use. They term this as the "efficiency-effectiveness tension", the trade-off between immediate task completion and genuine skill development. Other studies call for a more value-sensitive approach. Zhu et al. [90] demonstrated that algorithmic systems must balance multiple stakeholder values and trade-offs, while Lee et al. [41] showed how participatory frameworks can navigate equity-efficiency tensions. We adopt these value-sensitive approaches to examine self-directed language learning, where delegation decisions involve navigating multiple tensions without institutional scaffolding.

3 Study 1: Language Learners' Online Discussions on LLMs

As an initial observation of the learner-side practices in self-directed language learning, we aimed to understand language learners' perceptions of LLMs and their adoption patterns in language learning contexts. To this end, we conducted an exploratory analysis on the r/languagelearning subreddit [64]. The overall process is depicted in Figure 1.

3.1 Contexts and Settings: Reddit Community of Language Learners

Reddit is one of the largest online community platforms, comprising numerous topic-based subreddits, and as of July 2025, the r/languagelearning subreddit is ranked in the top 1%. Compared to other communities, such as YouTube, Reddit is unique in that its posts center on critical discourse, where users deliberate on LLM use, debate conflicting values, and make conscious decisions about adoption or rejection. Through a sweeping analysis of discussions on LLM usage in language learning, we aim to identify a set of tasks learners often delegate to LLMs and surface key considerations that complicate LLM adoption.

3.2 Data Sampling and Study Procedures

We collected public posts from January 2025 to June 2025, as the rapid improvements in LLM performance may easily render previous experiences with the technology outdated. The data dumps of all posts and their comments were provided by Project Arctic Shift [5]. Posts removed by moderators or the submitter were initially filtered out, as their content was no longer reliably accessible for analysis. Filtering the posts that contain at least one of the LLM-related keywords yielded 457 posts with 6483 comments. Note that our keywords included much broader terms, such as *"AI"*, since most community members did not differentiate between LLMs and general AI when discussing LLM usage, likely due to lesser public familiarity with the term *"LLM"*. While we acknowledge the distinction between LLM and general AI terminology, we respect the original phrasing in quotes and use the terms interchangeably where it reflects the community's usage, except where a distinction is necessary for clarity.

We further refined this pool by excluding those with more than 70% of their comments that did not contain any of the keywords. This was to prioritize posts that discuss LLM as a central topic. Posts without comments were still included in the analysis. We further excluded non-English posts, discussions regarding non-LLM AI, and posts selected by false keyword matches such as *"j'ai"*('I have' in French) matching *"AI"*. This resulted in a total of 191 posts(Table 1). Among the 1614 comments to these posts, 777(approx 48.1%) contained any of the keywords(Table 2). Before analysis, any identifying details, including the usernames, were removed or anonymized from both posts and comments. Any quote cited in this section is paraphrased to prevent identification of original sources through search engines.

3.3 Data Analysis

We adopt the reflexive thematic analysis (RTA) method demonstrated by Braun and Clarke [9, 10], a qualitative analysis method with theoretical flexibility and a constructionist focus. The first author initially familiarized themselves with the community and the data by browsing AI-related posts and community guidelines on AI usage. Removing non-English posts and false matches also involved a thorough reading of the collected data, which helped further familiarization. Next, the first author initially coded the data with a focus on tasks (RQ1) and rationales (RQ2) for the use and non-use of LLMs. Although the post-comment structure was preserved for analysis, the two types of text were merged and analyzed together with a shared code pool. The themes from the resulting codes were iteratively developed and defined through discussions between the authors.

4 Findings from Study 1: Online Community Discussions

To understand the considerations that shape learners' delegation decisions, we first established the common task space observed in LLM use cases. Then we examined the underlying dimensions along which community members evaluated LLM delegation. From online discussions, we identified three major value dimensions that guided these decisions: ACCURACY, INDEPENDENCE, and AUTHENTICITY. These dimensions capture how learners justified or rejected delegation, reflecting not just practical trade-offs but also the fundamental values they attached to the learning process itself.

4.1 Task Taxonomy for Language Learning with LLMs

We provide an overview of the tasks by category to sketch the decision space, drawing on subreddit discussions of which language learning tasks were appropriate for LLM use. To understand the tasks in the context of self-directed learning, we adopt Zimmerman's model of self-regulation in learning [92]. This cyclic model consists of three phases: forethought, performance, and self-reflection. To differentiate the externalized tasks from the internal phases, we derive the **planning** task from the forethought phase and the **evaluation** task from the self-reflection phase. We further subdivide the performance phase to capture distinct learning areas in language learning: **conceptual explanations**, **language input practice** (e.g., reading and listening), and **language output practice** (e.g., speaking and writing). Together, this taxonomy outlines

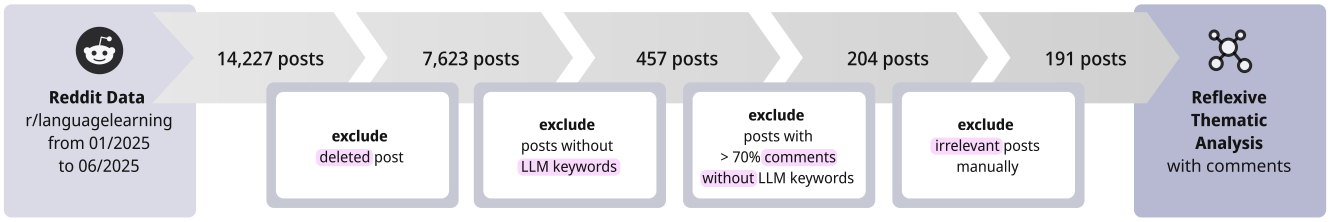


Figure 1: Study 1 process overview. The study was based on the Reddit posts and comments of r/languagelearning from January 2025 to June 2025. The data sampling process was designed to sample the posts that have LLMs as their main topic and resulted in 191 posts. These posts, along with their comments, were analyzed using the reflexive thematic analysis method.

the space of possible language learning with LLMs. The overview of the tasks can be found in Figure 2.

4.1.1 Planning. Though less common than other tasks, some learners shared their experiences using LLM assistance to manage learning content and execution that are not directly related to any of the previously discussed activities. One learner shared their prompt to generate and execute lessons tailored to specified learning needs and conditions, including neurodiversity supports. Another mentioned that they believed AI could help them decide which resource or learning method to use when there are too many options.

4.1.2 Concept Explanation. One direct application of LLMs is using them as language tutors to fill gaps in their understanding. Usage patterns included requesting full explanations of certain concepts or asking context-specific questions that arose while studying other outside materials. Topics ranged from fundamentals, such as vocabulary definitions and grammar rules, to more exploratory topics, including the cultural backgrounds behind language use.

4.1.3 Language Input Practice. LLM was often a solution to prepare resources for language input practice. Some learners generated entirely new materials, while others would employ LLMs to curate existing articles, books, videos, or songs. A hybrid approach involved rewriting and adjusting existing content based on individual preferences. This included adaptations made to improve out-of-date, overly formal, or inconsistent materials. In all cases, LLMs were favored for their ability to prepare for a custom level, topic, or learning focus.

4.1.4 Language Output Practice. Another popular method was to use the LLM as a practice partner for language output, such as speaking and writing. In practicing conversations, LLM was usually asked to assume the role of a native speaker to talk to. While traditionally done through text-based chat, the introduction of speech-based modes, including ChatGPT's advanced voice mode, has also made LLM-assisted speaking practice easily accessible. For writing exercises, LLMs were often used to generate writing prompt ideas. Learners would also come up with more engaging activities, such as role-playing games, text-based adventures, and guessing games.

4.1.5 Evaluation. The output practice using LLMs was often combined with personalized feedback from LLMs. Learners would ask for varying improvements in their language use, asking LLMs to point out grammatical errors, suggest alternative ways the natives would use in real life, and come up with more advanced words and expressions to further push their boundaries. Alternatively, the learners also requested an overall evaluation of their proficiency levels and an analysis of their weak points to help monitor their progress and determine their next learning focus.

4.2 Accuracy: Considerations in General Usages

As in other general tasks involving LLMs, the risk of inaccurate responses and hallucinations was the most serious concern of LLM usage in language learning. Community members noted that false information would be conveyed without transparency regarding accuracy or confidence levels. "It's like 85% solid, but the other 15% is just straight-up nonsense said with total confidence." They attributed this to the underlying mechanism: LLMs generate realistic answers without actually understanding the language. Users noted that it is particularly problematic in learning, as learners often lack the proficiency necessary to identify errors. For those who opposed LLM use, this was often the cause of their apprehension, which significantly contributed to their withdrawal. In language learning contexts, users also identified more subtle quality issues in LLM-generated content. While not grammatically incorrect, LLMs are reported to fail to accurately reflect actual language use. This led to concerns about the possibility of ending up "talking like an AI".

To cope with the limitations of their knowledge, learners often actively evaluated the quality of LLM output. One strategy was to test using their native language, which they could judge with confidence. Those who have access to teachers or native speakers often ask them to validate LLM outputs, adjusting their trust levels accordingly. The online community also provided support, as members actively shared their experiences of identifying errors that contradicted their prior knowledge or previous LLM outputs, thereby building a shared understanding of AI capabilities. They acknowledged the varied performance across target language, proficiency level, and model type, and actively sought information tailored to their individual needs.

As they make judgments on the reliability of LLMs, they find themselves in a trade-off: should they endure the risks of inaccurate information for the sake of efficiency in time and effort? Language learners find different points of balance depending on their tolerance levels. On one end, some users believed "AI could

Table 1: Number of Resulting Posts and Comments for Each Month

	JAN	FEB	MAR	APR	MAY	JUN	Total
Posts	14	12	25	28	55	57	**191**
Comments	81	15	182	186	579	571	**1614**
Total	95	27	207[1]	214	634[2]	628	**1805**

[1]March 16th: the community restriction on all AI-related posts was lifted [76]. This rule was meant to prevent the community from being flooded with duplicate questions and promotions of low-quality AI-based apps.

[2]April 28th: CEO of Duolingo, a popular language learning app, announced to go "AI first" [23]. This sparked active debates on using AI-based language learning apps.

Table 2: Number of Posts and Comments Containing Each Keyword

Categories	Keyword[1]	Posts	Comments	Total
LLM	LLM	8	59	67
	language model[2]	0	6	6
Related generic terms	AI	150	575	725
	IA[3]	3	2	5
	artificial intelligence	0	4	4
	chatbot	6	24	30
Major LLM-powered agents	ChatGPT	40	176	216
	GPT	18	41	59
	Gemini	7	26	33
	DeepSeek	2	6	8
	Claude	1	7	8
	Copilot	0	8	8
	Grok	0	7	7
	Perplexity	0	4	4
	LLaMA	0	1	1
	Qwen	0	0	0
	Bard	0	0	0
Any of the keywords		191	777	968

[1]Including their capitalized/plural forms and occurrences that are separated by symbols or numbers (e.g., GPT4, gpt-like)

[2]Catches "large language model" as well.

[3]Community alternative for AI used to avoid the automatic removal due to the community rule (see Table 1).

not be trusted with anything (related to learning)", and that they would "rather go straight to reliable sources than to cross-check AI outputs." On the other hand, some users actively used LLMs for all tasks, either because they believed the error rate was negligible or because they didn't expect perfection. Some claimed that with suitable, detailed prompts, such errors could be minimized and that the reported errors were rather a user issue. Others pointed out that every source has a risk of being wrong, as they would say, "humans make mistakes too," or "same way it's foolish to think everything posted on Reddit is accurate." Otherwise, they simply couldn't resist using it, even knowing it may be wrong, because they "find it too useful".

LLM tasks appear to exist along a perceived risk spectrum. For instance, LLMs were often thought of as unsuitable to ask for evaluation, since they will "tell you whatever they believe you want to hear" without actual judgment. Asking LLMs for detailed explanations was also considered very risky, as learning the wrong information could be critical. Simple questions, such as word definitions, were considered to be less likely to be wrong. Generating grammatical example sentences was thought to be a task that LLMs are good at. For practice partners, users didn't consider hallucinations as detrimental, because it was more important that they had somebody, or something, to talk to. One user advocated using LLMs for practice, saying, "learning language with a real person costs

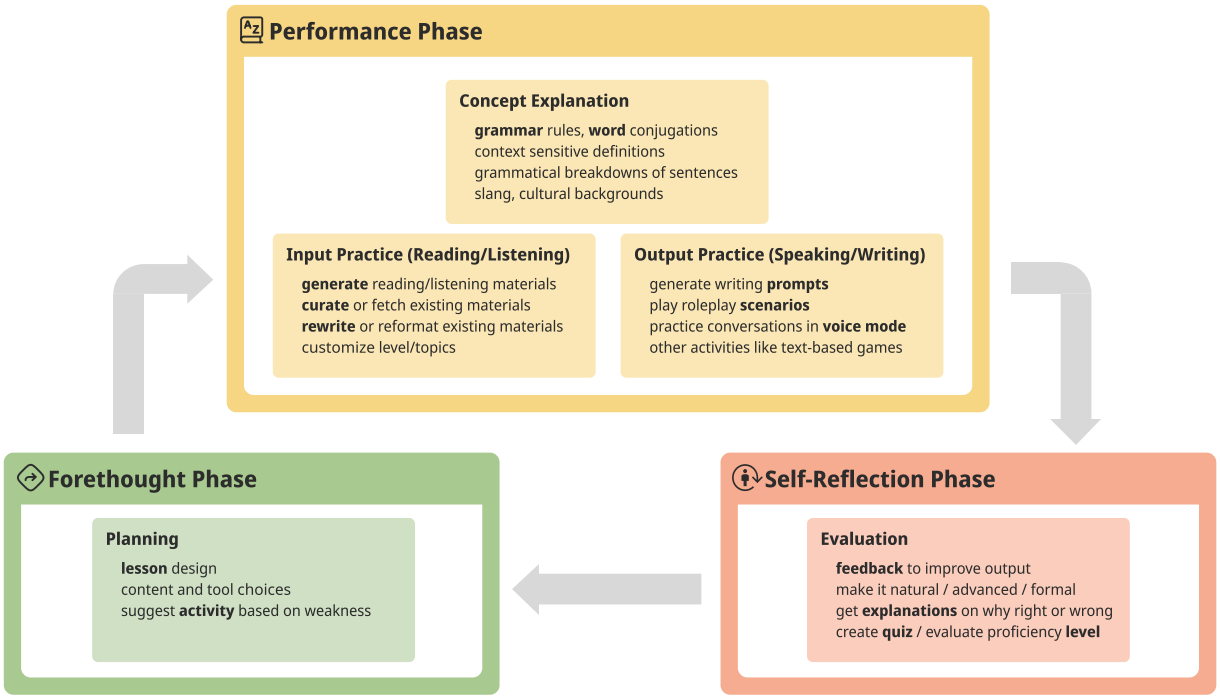


Figure 2: Results from Study 1: Task taxonomy for language learning with LLMs, based on Zimmerman's self-regulation model in learning. The taxonomy organizes language learning tasks into a three-phase cycle: forethought phase, performance phase, and self-reflection phase. The performance phase is further subdivided into three categories to capture distinct learning areas in language learning: conceptual explanations, language input practice, and language output practice. Each phase represents different types of learning activities that language learners discuss in relation to LLM usage.

a lot... let AI handle the filler work and use humans for the stuff that actually matters." More rigorous learners would limit LLMs to a tool for fetching human-made resources according to a custom criterion.

4.3 Independence: Considerations in Learning Context

However, independent task completion offers a distinct value in learning contexts. Because the goal is not to produce a specific output but to develop the skills required to build it on their own, learners should decide which task is essential for them to do independently. One of the common arguments was that the skills to answer the questions and prepare the required management work is not just an unnecessary overhead, but also a crucial skill required in learning. This included looking up dictionaries and inferring meanings from context, as shown in this comment: "You're also taking away essential abilities like finding verb forms, word definitions, working out what sentences mean, spotting known and unknown patterns. For what purpose? To understand the text?"

This criticism is not limited to specific tasks such as googling or searching in textbooks; it generalizes into a classical concern about cognitive offloading or outsourcing essential learning activities. There were multiple mentions of probable cognitive skill atrophy, including critical thinking and communication. One user confesses that they feel their language skills actually dwindled: "after 2023, as I've started using ChatGPT and even downloaded the app on my phone, I wasn't able to even remember these phrases (that I actually had memorized in the last years)". Others, more favorable to LLM use, view this as a repetitive argument that would be outdated, as it has been for previous technologies. One user simply puts, "I'm sure some cranky old person was also upset when Google Maps first came out."

Another classical concern is that because the learning process naturally demands substantial engagement, attempts to make it convenient risk being self-defeating. Simply put, what is easily gained is easily lost. According to these arguments, creating resources is a good exercise in itself, and it is not advisable to outsource it altogether to technology. One user replied to a request for an image-generation tool as a memory aid, that "the benefit comes from the actual work of creating them in your head, not just from the final mental picture." In deciding where to dedicate their efforts and what to delegate to LLMs, learners should consider what parts of learning they believe are central to their learning in terms of involved skills and mental dedication.

4.4 Authenticity: Considerations in Language Learning

In considering LLM assistance, language learning is unique in that its goal is to communicate with humans. Several Reddit users expressed discomfort with learning a human language through interaction with a non-human entity. For them, this is a paradoxical approach ("Why would anyone want a program to teach you how people actually talk?!"). They believe that AIs lack understanding of history, culture, and underlying emotions, which makes it nonsensical for them to teach idioms and phrases rich with cultural and emotional connotations. As conversation partners, they were often described as artificial and repetitive. Several users complained that they have to drive the conversation as LLMs do not take an active role, or that it is not engaging enough, knowing that LLM is "pretending to know or care". Even as tutors, some say it is less engaging or motivating compared to human teachers.

To others, the non-human aspect of LLM-infused learning is more of a liberation. The most common pain points in practicing conversation were, predictably, the lack of native speakers who can be patient with beginners and can provide feedback and explanations. Online learning communities and direct messaging posed risks of encountering disrespectful strangers, and it was also difficult to enforce a desired practice format. As one proponent of AI conversation said, "they either ignore me completely, insult me with small spelling mistakes, or just switch to English." Tutors were not as affordable, especially for lengthy practices and question-answering, and they were not as readily available as LLMs.

Social dynamics were one of the main "sparks" of human interaction, but also made learning stressful for novices. One common motivation for using LLMs instead was that they didn't want to bother native speakers with small, many, or even 'stupid' questions. While suitable levels of pressure may lead to better engagement, a fear of judgment or anxiety can negatively impact their willingness to communicate. Most learners agreed that direct communication with real people was irreplaceable, but proponents of LLM claimed to use it as a stepping stone before actually facing native speakers.

5 Study 2: Technology Probe on LLM Use in Language Learning

While Reddit discussions offered a broad view of how language learners perceive the usage of LLMs, they often lacked the situated context of actual learning activities and the detailed rationales and backgrounds required for an in-depth understanding of learners. Moreover, such discussions could not surface unconscious considerations that shaped delegation decisions. Therefore, we conducted a technology probe study that placed learners in a structured yet flexible learning environment to further observe the decision-making process.

5.1 Context and Settings: Probe Design

5.1.1 Design Goals. Our system is designed as scaffolding rather than active guidance to preserve learner autonomy in self-directed learning. The system is built around a familiar chat interface to minimize cognitive burden and preserve the natural environment where learners use LLMs. The system supports autonomous exploration of LLM integration without providing explicit instructions.

- **DG1: Metacognitive support for self-directed learning.** The system workflow should follow Zimmerman's model (forethought, performance, and self-reflection) [92] to facilitate deliberate decision-making about task delegation.
- **DG2: Scaffolding for LLM utilization.** The system should present the space of usage cases for experimentation, while reducing barriers to effective prompting.

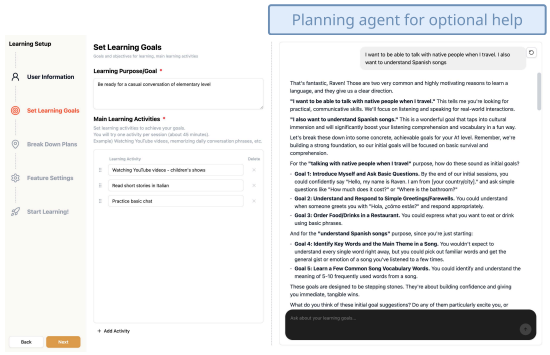
5.1.2 User Interface. To support DG1, the system comprises a **Setup Page** for the Forethought phase and a **Learning Page** for the Performance phase of Zimmerman's model. The Setup Page scaffolds the Forethought phase through three stages: **User Profiling** to specify language level and context, **Goal Setting** (Figure 3(A)) to define learning objectives and activities, and **Activity Breakdown** (Figure 3(B)) to decompose into sub-tasks. This phase prepares learners to be self-aware of their current status and goals, and facilitates deliberate decisions about LLM integration. Learners can optionally consult the LLM during goal setting and planning.

The Learning Page facilitates the Performance phase through a dual-panel interface that balances goal awareness with active learning engagement. The **Learning Context Panel** (Figure 3(C), Left) persistently displays the learning objectives and activity plans established during Setup, serving as a constant reminder that supports self-monitoring, a key component of Zimmerman's Performance phase. The **Chat Panel** (Figure 3(C), Right) features a chat session where learners engage with the system through an agent selection mechanism: either selecting automatic mode, where the system determines the appropriate agent, or directly choosing from five specialized agents(planning, input practice, concept explanation, output practice, evaluation).

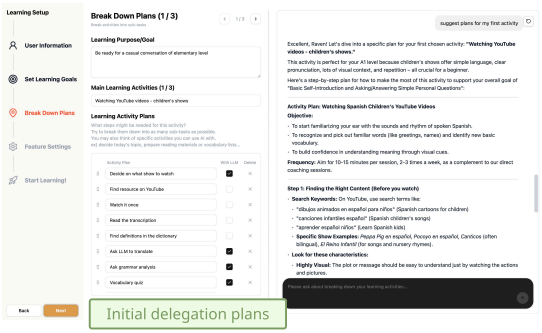
The probe also includes an **Agent Configuration** (Figure 3(D)), accessible from both the Setup and Learning Page. It allows learners to customize task-specific preferences and response styles, with changes applied instantly while preserving the learning context, thereby reducing the load of formulating requests.

5.1.3 Multi-agent Design. To achieve DG2, the system addresses a critical challenge: enabling learners to engage in SDL-aligned interactions with LLMs without requiring expertise in prompt engineering or complex instruction formulation. The system resolves this through a multi-agent architecture that decomposes common language learning interactions into specialized agent roles, allowing learners to access appropriate LLM support through intuitive agent selection rather than manual prompt construction. The system implements a root coordinator agent that orchestrates five specialized agents, analyzing learner requests and routing them to the appropriate agent while maintaining conversation context across agent transitions. Learners can either rely on this automatic routing mechanism or directly select a specific agent through the interface, preserving their autonomy in choosing support strategies.

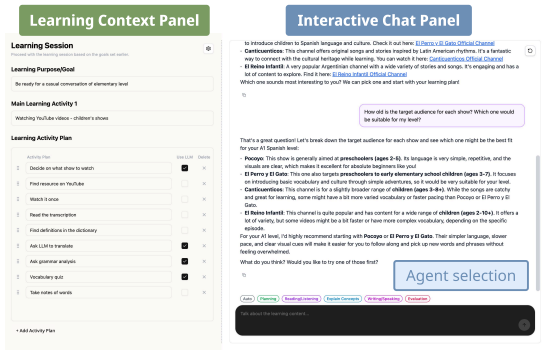
The five specialized agents, all powered by the Gemini-2.5-flash model, are equipped with distinct capabilities and tools tailored to their roles. The **planning agent** provides strategic guidance without additional tools, focusing on planning and decision-making support. The **input practice agent** integrates web search and file generation tools (.md, .csv, .docx, .pdf) to create and curate learning resources, while also accessing the evaluation agent for quality



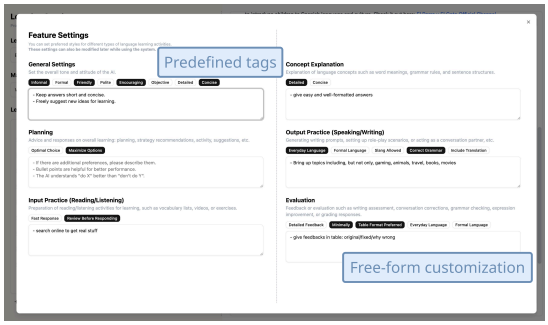
(A) Setup Page: Goal Setting Interface



(B) Setup Page: Activity Breakdown Interface



(C) Learning Page



(D) Agent Configuration

Figure 3: Probe system overview. (A) and (B) are from the Setup Page that the learners use to specify plans in the initial session. (C) is the Learning Page interface used in the learning session. (D) is the agent configuration accessible both in the Setup Page and the Learning Page.

validation. The **concept explanation agent** combines search capabilities with file generation to provide accurate grammatical and vocabulary clarifications with supporting documentation. The **output practice agent** facilitates interactive exercises and can invoke the feedback agent to assess learner performance during activities. The **evaluation agent** analyzes learner output and can consult the planning agent for strategic recommendations based on identified strengths and weaknesses. Each agent shares access to the learner's profile, goals, and activity plans established during Setup, enabling coherent support while preserving the specialized expertise that makes each interaction effective, without requiring learners to understand the underlying complexity of the prompt engineering.

5.2 Participant Sampling and Study Procedures

5.2.1 Participants. The participants were recruited from online community boards of a Korean university and included undergraduate, graduate, and alumni students. The screening survey required participants to self-report their target language, current proficiency, and LLM usage to maintain diversity in the participant pool. The final set of participants consisted of 13 learners of seven different languages, comprising seven beginners and eight intermediate learners. The participants' basic information is presented in Table 3. Each participant will be referred to by their participant ID, labeled with their target language. The study was reviewed and approved by the University ethics review board, and all participants provided written consent before participating. The participants were rewarded with 40,000 KRW (approx. 29 USD) in compensation for their time.

5.2.2 Initial Session. The initial session consisted of explaining the study logistics, interviewing participants about their past experience with learning and LLM, and planning learning activities to test on the system. This session was conducted either in person or remotely, depending on the participant's preference. After receiving a walkthrough of the study process and signing the consent form, participants were interviewed for 30 minutes about their past experiences with language learning and LLM utilization. These interviews were conducted to enable the researcher to gain an initial understanding of each participant's background and to help the participants themselves set a clearer picture of their learning goals, motivations, and status before planning their self-directed learning activities. The participants were then guided through the system's learning setup phase, as detailed in Section 5.1.2. Although the learning tasks do not have to be digital or text-based, participants

Table 3: Overview of Language Learning Tasks Assisted by LLMs

ID	Age	Gender	Target lang.	Curr. Lvl.[1]	Learning Purpose/Goal
CN1	21	F	Chinese	A1	For casual conversation
CN2	26	F	Chinese	B1	For casual conversation
CN3	24	M	Chinese	B1	For work
DE1	25	F	German	A1	As a hobby
EN1	29	M	English	B2	For work
EN2	26	M	English	A2	For job interview
ES1	20	F	Spanish	A2	As a hobby
FR1	25	M	French	A1	For research
FR2	24	F	French	A1	For job application
JP1	27	F	Japanese	B1	For media consumption
JP2	30	F	Japanese	B1	For career change
JP3	28	F	Japanese	B2	For job / media consumption
RU1	30	F	Russian	A1	As a hobby

[1]Self-reported language proficiency levels based on the Common European Framework of Reference for Languages (CEFR), ranging from A1 (beginner) to C2 (advanced).

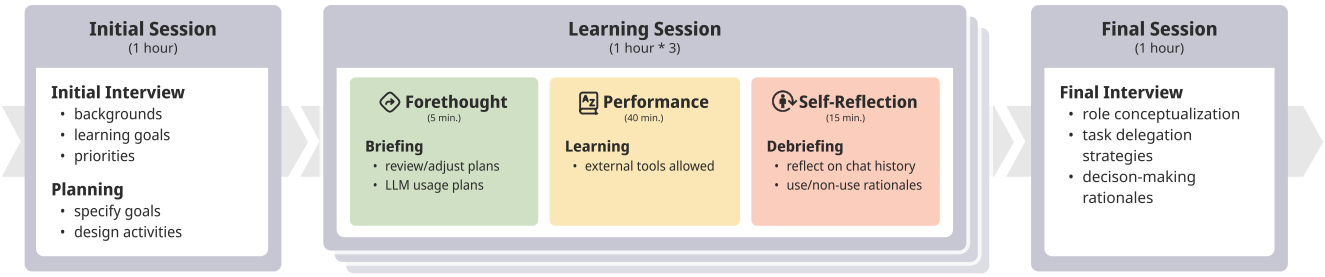


Figure 4: Study 2 process overview. The study consisted of three phases. The initial session included consent procedures, an interview about past language learning experiences, and a planning task where participants set goals and designed activities using the probe system. Each participant then completed three one-hour learning sessions, each comprising a brief forethought phase (5 minutes), a performance phase of LLM-assisted learning (40 minutes), and a self-reflection phase (15 minutes) to review LLM use and non-use rationales. The study concluded with a reflective session, a one-hour semi-structured interview.

were asked to focus on the tasks they would want LLM assistance with most. The initial session concluded with a guided tutorial on the system's features, including agent usage and customization.

5.2.3 Learning Session. Each participant had three learning sessions, one for each activity devised in the initial planning. Each session lasted an hour and was primarily conducted remotely in their typical learning environment, using their personal computers. Learning sessions were scheduled on separate days to control for participant fatigue. While the entire screen was shared with the researcher for observation, no intervention was made other than the initial instruction and resolving technical issues during learning. Participants were asked to use the system to conduct their learning activities, with the flexibility to modify their LLM usage strategies as needed. The initial plans, including their LLM usage strategies, were presented on the left side of the system and could be modified at any time. During the learning process, participants were encouraged to experiment with their delegation strategies, prompts, and agent settings. After 45 minutes of system usage, the participants were asked to debrief on LLM utilization, explain the rationales and perceptions of specific interactions, and discuss how they would modify the strategies, if any.

5.2.4 Reflective Session. The reflective session was an hour-long, semi-structured interview that examined 1) the participant's role conceptualization and task delegation strategies, and 2) their decision-making rationales, including their opinions on value conflicts surrounding ACCURACY, INDEPENDENCE, and AUTHENTICITY. The first author, who conducted all the previous sessions, designed individual follow-up questions based on each participant's past quotes and their interaction patterns.

5.3 Data Analysis

5.3.1 Log Analysis. The probe logged all learner interactions throughout the study, including message content, agent selections, and plan edits. Only learner responses were coded, while the preceding LLM outputs and the surrounding learning context were referenced to interpret the meaning of each entry. The log analysis followed a hybrid coding approach [24]. First, all responses were deductively classified according to the taxonomy of language learning tasks developed in Study 1: Planning, Input, Output, Explanation, and Evaluation. Second, within each category, we conducted inductive, reflexive thematic analysis (RTA) [9, 10] to identify the tasks learners used to interact with the LLM. Because individual responses often involved multiple functions, entries were allowed to receive multiple category codes. The first and second authors were familiarized with the full log data prior to the analysis. The initial codes were developed by the second author and then synthesized into subtasks and tasks through iterative discussions among the authors.

5.3.2 Interview Analysis. The interviews, briefings, and debriefings were fully recorded with participant consent and then transcribed using an automatic transcription tool. The debriefings and the final interview were the primary targets of analysis. The initial interview was designed to understand individual backgrounds and contexts, often containing opinions and experiences in general LLM use or learning. Initial interviews were coded separately and were used as supplementary information to better understand excerpts from later sessions. Briefings served as background context to interpret participants' approaches and logs, but were not systematically coded or treated as primary data for the results.

The transcriptions of the debriefings and the final interviews were analyzed using the RTA method [9, 10]. The first and second authors were familiarized with the full dataset prior to analysis. The initial codes were developed by the first author and synthesized into subthemes and themes through iterative discussions among the authors. Throughout the process, the analysis was triangulated with the initial interviews, briefings, chat logs, and manual observations made by the first author during the sessions.

6 Findings from Study 2: Technology Probe

Section 6.1 reports the log analysis of enacted tasks; Sections 6.2-6.4 present interview-based findings along the three value dimensions guiding delegation decisions, which were first revealed in Study 1; and Section 6.5 introduces an additional insight about prompting from the interviews.

6.1 Observed Tasks in LLM Usage

Through log analysis, we identified 19 distinct subtasks across five top-level categories of language learning activities in Study 1 (Planning, Explanation, Input, Output, and Evaluation). The occurrences are summarized in Table 4. Learners engaged with most of the five language learning tasks. Eleven learners used all five categories at least once during the three study sessions. The remaining two engaged in four tasks each. Some tasks were nearly universal. All participants requested explanations. Twelve learners generated new materials, and twelve also searched for existing resources. At the session level, these activities were also dominant. Explanation tasks appeared in 34 of 39 sessions, and material generation in 25.

Output practice, as defined in Study 1 as the production of the target language, was less common and narrower in scope. It appeared only when learners engaged in role-play conversations or writing tasks. These activities were observed in 11 participants across 19 sessions, whereas input practice was observed in all participants across 33 sessions. In language learning, output practice is generally considered more effortful than input [71], and our study similarly showed that learners engaged less in output practice compared to input activities.

Evaluation tasks showed greater diversity in both how learners reflected and what they reflected on. Participants reviewed LLM outputs, occasionally their own process, but more often asked the LLM to provide feedback, generate practice problems, or check proficiency. Reflection, therefore, took multiple forms, but it was more often mediated through the LLM than carried out as explicit self-assessment. Overall, the log analysis confirmed that the task taxonomy from Study 1 captured learners' practices well.

Table 4: Sub-tasks observed in the probe study logs, grouped by task. The numbers indicate how many participants (out of 13) engaged in each sub-task at least once.

Task	Sub-task	Count
Planning	Request Suggestions	6
	Accept Suggestions	3
	Declare Plan	5
	Configure LLM	7
	Choose Content	5
Explanation	Ask Questions	13
Input	Generate Materials	12
	Search Resources	12
	Generate Summary	6
	Answer Questions	7
	Demonstrate Understanding	5
Output	Conversation	8
	Writing	5
Evaluation	Request Feedback	7
	Request Practice Problems	9
	Request Proficiency Assessment	6
	Self-assess Proficiency	4
	Reflect on Learning Process	4
	Evaluate LLM Output	7
Total participants (N)		13

Note: Each count indicates the number of unique participants who engaged in the sub-task at least once during Study 2. Multiple codes could be assigned to a single learner response if it reflected more than one sub-task.

6.2 Accuracy: Rationalizing by Individual Contexts

Probe participants provided more detailed rationales of accuracy assessments based on their learning goals, preferences, and current proficiency. All participants were aware of the possibilities of hallucinations while using LLMs, either from previous experiences in language-related tasks or from other domains in which they use LLMs. While some believed this risk could be mitigated through specific prompting, most felt limited in controlling output accuracy. Managing this "issue of the model itself (CN1)" took the form of a risk management problem. The participants would assess the risks, decide the extent to which they could tolerate them, and determine the optimal areas of use.

Unable to directly judge accuracy due to limited proficiency, participants relied on indirect assessments: past hallucination experiences in other domains, interaction quality in other languages, or perceived task difficulty. This would then be connected to the perceived difficulty of the task. Tasks that involve basic-level language, easily available information (e.g., online dictionaries), or verifiable outcomes were classified as inherently low-risk because they were considered well-represented in the training data. On the other hand, ambiguous grammar rules or subjective tasks related to style or appropriateness were considered higher risk and required more scrutiny.

While participants in the probe study were well aware of the risks of inaccuracy, they differed most from Reddit members in their willingness to tolerate errors. Casual learners constructed a pragmatic reasoning where communication effectiveness was prioritized over precision. They rationalized their error tolerance by the fact that minor mistakes did not impede their core goal of communication, making additional verification an inefficient use of limited time and effort. Learners with more serious motivations either thought the risk was negligible at their current level of proficiency or that they would eventually correct themselves as they learned more. The participants reasoned that they would validate the outputs in more serious use cases, as with JP1, who used a dedicated translator to cross-check LLM outputs before sending messages in Japanese, but considered learning scenarios as a low-stakes situation. DE1 associated her tolerance with learning preferences. "Right now I just read through it once or twice, maybe three times, say it out loud, and go 'oh okay, there's this rule' and move on. ... analyzing everything precisely and memorizing it all before moving forward, then for that kind of user it would probably be more critical."

In the initial interviews, most of the participants shared their experiences of hallucinations and how it was necessary to cross-check in other domains. Six of them mentioned that they should also cross-check for language learning. Despite acknowledging the importance of verification, participants rarely cross-checked LLM outputs in learning sessions, revealing a reasoning gap. In practice, cross-validation was a conditional process that was triggered by a clear contradiction with their prior knowledge or a previous LLM-generated explanation. Participants would often instead blindly trust LLM outputs because they cannot verify their accuracy (CN1, FR1, JP3, RU1), or because verifying is itself a significant effort (CN1, CN2). Users like CN3 and DE1 were also deterred by the tool switch's friction. The unresolved risk remains as a doubt that underlies the learning process (EN1, EN2, JP3). The exceptional case was where the participant used an external resource as a learning material. This led to spontaneous cross-checking and, as a result, much less anxiety about being wrong.

6.3 Independence: Maintaining, Enabling, or Overwhelming

Regarding the independence issue in using LLMs for language learning, participants were less concerned with cognitive offloading or overreliance than Reddit community users, and instead viewed LLM delegation as a resource-allocation problem. Many learners described their language learning as a side project or hobby, and they had limited time and cognitive resources available for learning. The participants were generally confident in their abilities to maintain efficient learning, including JP3 and RU1, who explicitly expressed that they were "tempted" to use LLMs for core learning activities. EN2, ES1, and JP2 thought that LLM performance was still too limited to displace their roles as learners. Conversely, CN3, envisioning that some language skills would be fully replaced by LLMs, believed it was enough to supplement his shortcomings with AI assistance.

The participants employed multiple boundary maintenance strategies to preserve their areas of effort rather than relying solely on LLMs. JP1 and RU1 retained their original learning approaches (textbook and Duolingo app, respectively) and integrated the LLM system for gap filling. Active recording strategies found in CN1, JP2, FR2 with different mediums reflected beliefs that physical engagement and active reorganization enhanced retention. Multiple participants experienced LLMs crossing their intended boundaries to give unwanted help. This included translations provided in advance, pronunciation notes that made vocabulary quizzes too easy, and sample sentences that gave away answers to writing exercises. External tools were not only a means of cross-checking but also served as intentional limitations on LLM support. CN1 and FR2 used online dictionaries to look up words and tried composing their own sentences. While many participants thought it was important to practice input or output on their own before asking the LLM for assistance, this was at times overwhelming for beginners. For instance, FR1 requested a full translation into French because he felt he needed more scaffolding before actual writing practice.

Participants experienced a reversal of cognitive burden: rather than fearing LLM dependency, they were overwhelmed by LLMs' dependency on their self-direction. The learners appreciated the LLM's potential to customize and control the learning process, actively deciding what to learn next, organizing various learning procedures, and adjusting the methods, frequency, and scrutiny of evaluations. Conversely, many of them expressed concerns that using LLMs in language learning required more SDL skills than traditional methods. As highly versatile tools, LLMs impose a greater burden of design, monitoring, and reflection. Traditional materials, such as textbooks, have a well-laid curriculum developed by human experts and were often trusted as a convenient and efficient approach to learning. Six participants noted that LLMs must be prompted to teach, making it difficult to discover new concepts they are not yet aware of.

While some participants discovered unexpected learning topics through LLM responses, this often led them to deviate from their original plans. Four were concerned about this unlimited freedom, or lack of structure, in LLM interaction that may make learners lose focus or skip necessary parts. Especially for concept explanation, structured alternatives would function as safety nets, providing coverage and a constructive constraint that supports concentration. For input or output practice activities, the versatility of LLMs was valued as adaptivity to various learning needs, both for generated and external content. Hence, the participants were more skeptical of LLM usage at the beginner level, believing it was more important to learn the basic concepts and structures than to practice actual language use.

6.4 Authenticity: Simulating Teachers and Learning Partners

The participants often compared LLMs with human support in language learning and considered whether they were a valid substitute for teachers or practice partners. They shared the popular opinions of Reddit on both sides of the debate: on how human imperfections, rapport, and sincerity make conversations authentic and engaging, and on how non-human LLMs could be an emotionally safer ground for beginners without social complications. Beyond the use of conversation practices, our participants tended to be more forgiving of unauthentic materials. While they recognized that reading materials were more generic or that there might be a gap with real-world language use, many of them thought authenticity was a low priority in the learning process. Some participants even thought the generic expressions would be ideal for beginners, describing them as "textbook-like" and therefore easier to learn.

In their interaction patterns, they showed distinct ways of treating LLMs. Some of them provided minimal prompts to achieve the desired output. This type of user often thought emotional LLM responses were performative rather than genuine. They chose to optimize for pragmatic content as CN1 would explicitly prompt: "No, don't try to encourage me, just give me more example sentences." When some of these users provided subjective inputs, they described these prompts as feedback to steer the AI towards their desired behavior. A few others deliberately sought emotional exchanges with the LLM, including comments on learning activities and playful interchanges. FR1 would keep the LLM updated with his external activities ("Until now, I tried focusing on Gargamel's lines in the video and how he pronounces them") and shared subjective remarks ("I laughed a lot mimicking Gargamel's lines"). ES1 asked for LLM's opinions on Spanish songs she studied, and told jokes about how it pretended to understand love songs ("But you are an AI, you don't have any girlfriend/boyfriend lol sorry").

When the interviewer asked about these incidents, many described them as habitual behaviors in their social interactions. While participants were aware of the non-human nature of LLMs, they still felt them as humans to some extent. JP2 was one of the users who provided only the necessary inputs and customized the given LLM to exclude unnecessary comments, yet still found the conversation format to be very engaging. JP3 still felt a certain extent of social anxiety when talking to LLMs using voice mode, even though she knew it was not a human. EN2 explicitly stated that he did not want any emotional support from LLMs and often thought of it as an overstatement, but still admitted that, "I was like, was it that good? ... I kind of got fired up too, and I think I was able to focus better."

Interestingly, some learners reasoned that this seemingly redundant interaction may help improve the learning experience. ES1 shared that "getting it to talk more" helped get more information, as in the episode where she got some interesting backstory of the song while chatting about the lyrics. FR1 explained that his additional remarks helped relax the atmosphere and made him ask more questions. JP1 described her feedback to steer LLM behavior as a process of "creating a bond with an account."

6.5 Balancing Prompting Efforts

Prompting effort emerged as the meta-constraint shaping all delegation decisions. As some Reddit users did to dismiss criticisms of LLMs, the participants often attributed their disappointments to the quality of their prompts. Participants recognized that better accuracy, independence, and authenticity were theoretically achievable through strategic prompting, but felt that effort-quality trade-offs forced them to compromise in their actual practices. Although our system did not provide specific prompts to avoid leading, several participants suggested providing a pre-made prompt set for learners to use.

7 Discussion

Self-directed language learners were actively making delegation decisions based on three major considerations: accuracy, independence, and authenticity. However, with general LLM services without pedagogical guardrails, they may adopt suboptimal strategies that undermine their own learning goals.

Observing SRL of university students, Foerst et al. identified two distinct causes for this inefficacy in strategies [25]. First, they can determine the optimal strategy, but do not adhere to it. Second, they select suboptimal strategies because they lack sufficient pedagogical knowledge. Our participants gave direct reports for the first obstacle, which we named the *execution challenge*. Some of their accounts also implied the second obstacle, which we will refer to as the *selection challenge*. After discussing each in detail, we provide design implications to address these two obstacles in the context of the three considerations.

7.1 Challenges in Delegation Strategy Execution

Our participants reported difficulties or reluctance in implementing their stated intentions when integrating LLMs into language learning. This pattern echoes documented phenomena in self-regulated learning research: Foerst et al. found that university students correctly identified beneficial SRL strategies 87–95% of the time, yet failed to implement these strategies 22–34% of the time [25]. Understanding the mechanisms behind these execution failures is essential for designing systems that support learners in maintaining their own strategic intentions.

The reasoning gap of cross-checking LLM accuracy. The most striking execution challenge emerged around accuracy verification. In initial interviews, most participants acknowledged the

importance of cross-checking LLM outputs, drawing on their experiences with hallucinations in other domains. Six explicitly stated they should verify information when learning languages. Yet during learning sessions, verification was rare, triggered only when outputs clearly contradicted prior knowledge or previous LLM explanations. This reasoning gap, in which learners articulate sound strategies but fail to execute them, is a core challenge for LLM-assisted learning.

Several mechanisms explain this gap. First, verification imposes substantial effort costs. Participants described cross-validation as "in itself a significant effort" (CN1, CN2), and several were "deterred by the friction of the tool switch" (CN3, DE1). This aligns with Vasconcelos et al.'s cost-benefit framework, which reframes over-reliance not as an inevitable cognitive bias but as the outcome of rational effort-utility trade-offs [75]. When verification costs exceed perceived benefits, learners strategically choose to trust, even when they know verification would be prudent.

In addition, our participants often described language learning as a hobby or a side project, and were limited in time and cognitive resources. Under such conditions, "satisficing" (accepting a "good enough" outcome rather than maximizing) becomes a rational strategy. Participants reasoned that minor errors would not impede their core communication goals or long-term learning outcomes, so additional verification would be an inefficient use of limited resources.

Violation of boundaries and INDEPENDENCE. The learners also struggled to maintain their intended boundaries around independent effort. Multiple participants experienced LLMs crossing their delegation boundaries to provide unwanted assistance: translations offered before learners attempted their own, pronunciation notations that made vocabulary quizzes substantially easier, and sample sentences that revealed answers to writing exercises. These violations undermined the productive struggle essential to skill development and eliminated the cognitive engagement that produces learning. However, the participants actively sought to mitigate these risks by switching to alternative tools they could better control, such as online dictionaries.

Some participants also explicitly described being "tempted" to use LLMs for tasks they had initially designated as their own responsibility. This temptation persisted despite their stated commitment to independent practice, suggesting that in-the-moment convenience may override original intentions in real-world settings. Notably, participants generally expressed confidence in their ability to maintain boundaries—yet this confidence may be unwarranted. A previous study on writing found that survey respondents showed a comparatively lower level of self-monitoring, whereas interviewees believed they had critically reflected on their learning process [78]. Students who used AI scaffolding for peer feedback were also reported to be unable to replicate their skills without AI assistance [19]. Since evaluations on learning outcomes were outside our scope, it remains unclear whether self-directed learners actually maintain a stronger commitment to their goals or whether they overestimate their ability to do so.

7.2 Challenges in Delegation Strategy Selection

Execution challenges arise when learners know what they should do but fail to follow through. Selection challenges, by contrast, arise when learners lack the knowledge to make informed delegation decisions in the first place. This distinction matters for design: execution failures call for commitment devices and friction reduction, while selection failures require scaffolding that helps learners understand what approaches are appropriate for their situation. Our findings reveal that self-directed language learners face substantial selection challenges that compound the execution difficulties discussed above.

Limited capacity for direct ACCURACY assessment. The majority of participants expressed that they were incapable of direct judgment of LLM ACCURACY, due to their limited proficiency in the target language. This indicates the fundamental barrier unique to learning: the very knowledge required for such evaluation is what they are trying to acquire. Lai et al.'s work on conditional delegation highlights the problem: effective human-AI collaboration requires the ability to define "trustworthy regions" where AI outputs can be relied upon [40].

Since our participants lack this capacity, they make delegation decisions based on alternative evaluations. Some of them use available resources to validate, like native speakers or teachers. Others without such resources resort to indirect assessments: perceived task difficulty (i.e., questions on objective truth), past experiences in other domains (i.e., hallucinations in report writing), or folk theories about LLM capabilities (i.e., better with easily searchable definitions). However, learners have no systematic way to distinguish helpful intuitions from misleading ones, and the same content becomes inadequate as learners progress. Without guidance on LLM usage, learners must rely on trial and error to discover what works, and may never discover their previous mistakes.

Burden of INDEPENDENT instructional design. Opposed to the widespread concerns of overreliance, participants were often overwhelmed by the LLM's dependency on their self-direction. Traditional learning materials such as textbooks, courses, and apps embed curricular decisions made by experts. They provide structure, sequence, and content developmentally, and ensure coverage of essential topics. LLMs, by contrast, offer unlimited flexibility without inherent structure, placing the full burden of instructional design on learners themselves. Li et al. additionally point out that learners may struggle to leverage the full range of LLM affordances [43].

This finding invites comparison with intelligent tutoring systems (ITS), which maintain student models, adaptively sequence content, and ensure coverage of essential material [74]. In contrast, the flexibility of LLMs is especially of value for self-directed learners with varying needs. One promising direction is shared control: Corbalan et al. demonstrated benefits when an instructional agent selects a subset of tasks based on the learner's performance from which the learner makes the final decision [17]. Such approaches preserve learner agency while reducing design burden. However, this control still requires effective use of self-directed learning skills [18].

AUTHENTICITY and transferability left uncertain. AUTHENTICITY presented a different selection challenge than ACCURACY or INDEPENDENCE. Learners did not know what level of AUTHENTICITY

was appropriate for their proficiency. But unlike ACCURACY and INDEPENDENCE, which were conceptualized as trade-offs, they recognized the liberating values of a non-human language partner and regarded it more as a choice, depending on individual preferences. This liberation from social pressure is well documented: multiple studies show that AI chatbots reduce foreign language speaking anxiety and increase willingness to communicate [79, 82]. Commercial platforms explicitly market LLM-based chatbot practice as "a stepping stone to help you build up the confidence to speak your target language in the real world" [57]. Yet the critical question remains empirically unresolved: Does reduced anxiety with AI transfer to reduced anxiety with humans? A recent paper notes that research remains limited on whether chatbot-supported confidence persists outside classroom contexts [22].

7.3 Supporting Learners for Strategic SDL

Making validation accessible where it matters. (ACCURACY/Execution) As learners are conscious of the risk of inaccurate responses from LLMs, they adjust their tolerance levels according to individual goals and circumstances. This results in different learning trajectories: some avoid using LLMs entirely, rather than investing effort in validation. In contrast, others trust LLM outputs even in areas where they cannot directly assess credibility. Both groups find that the validation burden outweighs the benefits, highlighting a key design challenge. Rather than demanding users for constant scrutiny, systems could adopt a more satisficing behavior. But learners cannot identify which outputs need validation with their limited understanding. This suggests embedding pedagogical knowledge about verification priorities: which grammar patterns are foundational, which errors compound. Concentrating low-friction verification at these critical points could reduce tool-switching costs where ACCURACY matters most, paralleling how ITS encodes expert curriculum decisions.

Improving transparency for uncertainty. (ACCURACY/Selection) Learners struggle to correctly assess the credibility of LLM responses, which in turn increases the perceived risk, as they cannot detect potential errors or allocate validation efforts to where it is needed the most. This resonates with calls for greater explainability to allow learners to judge trustworthiness themselves [36]. While model confidence is known to affect user perception of ACCURACY [63] and their self-confidence [45], abstract metrics may be difficult for novices to interpret. Our participants made intuitive judgments about model reliability but lacked a systematic way to validate these theories. Future work is needed to characterize LLM trustworthiness across language learning tasks: where outputs are reliable, where regional variation complicates ACCURACY, and where errors are common. Surfacing this domain-specific knowledge could help learners refine their existing mental models and make more informed delegation decisions.

Facilitating conscious delegation decisions. (INDEPENDENCE/Execution) The learners were mostly confident in their own abilities to construct delegation boundaries. Despite concerns in the broader domain of learning [19, 88], learners reasoned that their language learning was self-motivated and that they would therefore not intentionally offload necessary tasks. Such over-delegation was instead attributed to LLM-side violations, inefficient prompting, or "temptations" to delegate what they originally thought was their own work. The language of temptation suggests a self-control problem, and our participants solved it not through willpower but through improvised commitment devices [11]: switching to dictionaries, keeping pen and paper, using external resources that structurally limited LLM involvement. In-the-moment delegation decisions, which are the default for general-purpose LLMs, are vulnerable to cognitive load and immediate convenience. Configuring conditional delegation [40] in advance may reinforce commitment for users, but pre-commitment mechanisms must therefore be low-friction, embedded in system defaults rather than requiring active setup.

Supporting adaptation of delegation boundaries. (INDEPENDENCE/Selection) Delegation boundaries are subject to adaptation across learning steps, situations, and content types. Some instances are part of the general strategy (e.g., trying by themselves first, then comparing with LLM's version) or would be a result of improvisation (e.g., adjusting the level of support depending on perceived difficulty). However, learners often cannot judge what adaptations are appropriate for their current proficiency. Shared control offers a promising direction: systems could propose boundary adjustments based on learner performance along with theory-informed rationales, with learners making final decisions [17]. This preserves agency while reducing the design burden that overwhelmed our participants. Systematic boundary adaptation settings may serve as an adaptive scaffolding, which is known to have positive effects on learning gains and SDL processes [46, 59, 80]

Degrees to simulating authentic language use. (AUTHENTICITY/Selection) There were multiple views on the ideal extent of simulating human conversations, implying the potential value of controlling the conversation style and format. While it was a matter of preference for some learners, other opinions were proficiency-dependent. The generated reading materials were considered generic and beginner-friendly, like textbooks. In contrast, real-time verbal conversation was challenging even with LLMs. These observations suggest a folk theory of AUTHENTICITY progression—from textbook-like text to chat to voice to human—that merits validation. While chatbot practice reduces speaking anxiety and increases willingness to communicate [79, 82], transfer to human interaction remains unexamined [22]. Moreover, idiodynamic research reveals distinct anxiety-willingness patterns across individuals and contexts [50], suggesting this progression may be preference-dependent. Design should support exploration across AUTHENTICITY levels rather than provide a fixed path. However, anthropomorphism of LLMs has its harms and risks, such as deception and misplaced trust [3, 16, 38]. A thorough investigation is needed on the effects and appropriateness of anthropomorphism in language learning contexts [21, 33, 66].

7.4 Limitations and Future Directions

Our research builds upon the framework of self-directed learning, with both our system and study specifically designed for deliberate delegation decisions. While this approach reflects the ideal procedure of SDL, it may introduce post-hoc rationalization, thereby obscuring the actual decision-making factors. Real-life LLM integration may also be influenced by unconscious or spontaneous

decision-making, which indicates a potentially valuable research area.

Our participant sample, aged 20-30 and drawn from a Korean university, may reflect specific cultural and educational contexts. While we believe the three considerations represent robust underlying rationales, the specific strategies they produce may vary with factors such as L1-L2 typological distance and prior multilingual experience, inviting further research on different populations. Additionally, while our taxonomy distinguishes broad learning areas (input, concept explanation, output), our analytical focus was on how learners reason about delegation rather than mapping task-specific patterns. Future work could systematically investigate how delegation strategies vary across specific learning types—such as vocabulary acquisition versus grammar learning—and proficiency levels. A long-term study on strategy evolution and learning outcomes is needed to validate whether learners' strategies lead to proficiency gains and to test system designs informed by our framework. We also note that our system lacked a speech module and was limited in capturing the additional dimension of multimodality.

8 Conclusion

This study reveals how self-directed language learners navigate LLM integration by simultaneously considering ACCURACY, INDEPENDENCE, and AUTHENTICITY; each dimension is shaped by individual learning contexts, goals, and histories. We found that learners actively construct delegation boundaries while managing multiple tensions: 1) tolerating ACCURACY risks based on task criticality and personal learning goals, 2) preserving learning effort while seeking efficiency, and 3) negotiating the paradox of seeking human-like language practice from non-human entities. These decisions are complicated by selection challenges, where learners lack knowledge to choose appropriate strategies, and execution challenges, where learners fail to follow through on stated intentions—as seen in the reasoning gap between cross-checking intentions and actual behavior. These findings suggest that effective AI-assisted learning systems must support conscious boundary-making, provide transparent uncertainty indicators, facilitate easy validation, and adapt to heterogeneous learner needs—ultimately enabling learners to craft personalized delegation strategies that preserve both learning efficacy and learner agency.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C200520911), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], and by the SNU-Global Excellence Research Center establishment project. The ICT at Seoul National University provided research facilities for this study.

References

[1] Safaa M Abdelhalim. 2024. Using ChatGPT to promote research competency: English as a Foreign Language undergraduates' perceptions and practices across varied metacognitive awareness levels. *Journal of Computer Assisted Learning* 40, 3 (2024), 1261–1275. doi:10.1111/jcal.12948

[2] O. S. Adewale, O. C. Agbonifo, E. O. Ibam, A. I. Makinde, O. K. Boyinbode, B. A. Ojokoh, O. Olabode, M. S. Omirin, and S. O. Olatunji. 2024. Design of a personalised adaptive ubiquitous learning system. *Interactive Learning Environments* 32, 1 (2024), 208–228. arXiv:https://doi.org/10.1080/10494820.2022.2084114 doi:10.1080/10494820.2022.2084114

[3] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All Too Human? Mapping and Mitigating the Risk from Anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 13–26. doi:10.1609/aies.v7i1.31613

[4] John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. 2004. An integrated theory of the mind. *Psychological review* 111, 4 (2004), 1036.

[5] ArthurHeitmann. 2025. *Project Arctic Shift.* https://github.com/ArthurHeitmann/arctic_shift Accessed: 2025-06-09.

[6] Seyyed Kazem Banihashem, Nafiseh Taghizadeh Kerman, Omid Noroozi, Jewoong Moon, and Hendrik Drachsler. 2024. Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education* 21, 1 (2024), 23. doi:10.1186/s41239-024-00455-4

[7] Azzeddine Boudouaia, Samia Mouas, and Bochra Kouider. 2024. A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Journal of Educational Computing Research* 62, 6 (2024), 1289–1317. doi:10.1177/07356331241247465

[8] Stefanie L. Boyer, Diane R. Edmondson, Andrew B. Artis, and David Fleming. 2014. Self-Directed Learning: A Tool for Lifelong Learning. *Journal of Marketing Education* 36, 1 (2014), 20–32. arXiv:https://doi.org/10.1177/0273475313494010 doi:10.1177/0273475313494010

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa doi:10.1191/1478088706qp063oa

[10] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. arXiv:https://doi.org/10.1080/2159676X.2019.1628806 doi:10.1080/2159676X.2019.1628806

[11] Gharad Bryan, Dean Karlan, and Scott Nelson. 2010. Commitment devices. *Annu. Rev. Econ.* 2, 1 (2010), 671–698. doi:10.1146/annurev.economics.102308.124324

[12] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[13] Philip C Candy. 1991. *Self-Direction for Lifelong Learning. A Comprehensive Guide to Theory and Practice.* ERIC.

[14] Wenli-Li Chang and Jerry Chih-Yuan Sun. 2024. Evaluating AI's impact on self-regulated language learning: A systematic review. *System* 126 (2024), 103484. doi:10.1016/j.system.2024.103484

[15] X. L. Chen, D. Zou, H. R. Xie, and F. Su. 2021. Twenty-five years of computer-assisted language learning: A topic modeling analysis. *Language Learning & Technology* 25, 3 (2021), 151–185. doi:10.64152/10125/73454

[16] Jennifer Chien and David Danks. 2024. Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 933–946. doi:10.1145/3630106.3658946

[17] Gemma Corbalan, Liesbeth Kester, and Jeroen J. G. Van Merriënboer. 2006. Towards a personalized task selection model with shared instructional control. *Instructional Science* 34, 5 (01 Sep 2006), 399–422. doi:10.1007/s11251-005-5774-2

[18] Gemma Corbalan, Jeroen JG van Merriënboer, and Wendy Kicken. 2010. Shared control over task selection: A way out of the self-directed learning paradox? *Technology, Instruction, Cognition & Learning* 8, 2 (2010).

[19] Ali Darvishi, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. 2024. Impact of AI assistance on student agency. *Computers & Education* 210 (2024), 104967. doi:10.1016/j.compedu.2023.104967

[20] Jesus De La Fuente, Paul Sander, Douglas F Kauffman, and Meryem Yilmaz Soylu. 2020. Differential effects of self-vs. external-regulation on learning approaches, academic achievement, and satisfaction in undergraduate students. *Frontiers in Psychology* 11 (2020), 543884. doi:10.3389/fpsyg.2020.543884

[21] Alicia DeVrio, Myra Cheng, Lisa Egede, Alexandra Olteanu, and Su Lin Blodgett. 2025. A Taxonomy of Linguistic Expressions That Contribute To Anthropomorphism of Language Technologies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 430, 18 pages. doi:10.1145/3706598.3714038

[22] Dongliang Ding and Ahmad Muhyiddin B Yusof. 2025. Investigating the role of AI-powered conversation bots in enhancing L2 speaking skills and reducing speaking anxiety: a mixed methods study. *Humanities and Social Sciences Communications* 12, 1 (01 Aug 2025), 1223. doi:10.1057/s41599-025-05550-z

[23] Duolingo. 2025. Below is an all-hands email from our CEO, Luis von Ahn – we are going to be AI-first. https://www.linkedin.com/posts/duolingo_below-is-an-all-hands-email-from-our-activity-7322560534824865792-l9vh?utm_source=

share&utm_medium=member_desktop&rcm=ACoAAFoCi3sB_PvT7nP1_q_rvmipAyMKQyJCRuk. [Accessed 12-08-2025].

[24] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5 (03 2006), 1–11. doi:10.1177/160940690600500107

[25] Nora M. Foerst, Julia Klug, Gregor Jöstl, Christiane Spiel, and Barbara Schober. 2017. Knowledge vs. Action: Discrepancies in University Students' Knowledge about and Self-Reported Use of Self-Regulated Learning Strategies. *Frontiers in Psychology* Volume 8 - 2017 (2017). doi:10.3389/fpsyg.2017.01288

[26] Dennis Foung, Linda Lin, and Julia Chen. 2024. Reinventing assessments with ChatGPT and other online tools: Opportunities for GenAI-empowered assessment practices. *Computers and Education: Artificial Intelligence* 6 (2024), 100250.

[27] D Randy Garrison. 1997. Self-directed learning: Toward a comprehensive model. *Adult education quarterly* 48, 1 (1997), 18–33. doi:10.1177/074171369704800103

[28] Wentao Ge, Yuqing Sun, Ziyan Wang, Haoyue Zheng, Weiyang He, Piaohong Wang, Qianyu Zhu, and Benyou Wang. 2025. SRLAgent: Enhancing Self-Regulated Learning Skills through Gamification and LLM Assistance. arXiv:2506.09968 [cs.HC] https://arxiv.org/abs/2506.09968

[29] Mohammad Ghafouri. 2024. ChatGPT: The catalyst for teacher-student rapport and grit development in L2 class. *System* 120 (2024), 103209. doi:10.1016/j.system.2023.103209

[30] Mohammad Ghafouri, Jaleh Hassaskhah, and Amir Mahdavi-Zafarghandi. 2024. From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *Language Teaching Research* (2024), 13621688241239764. doi:10.1177/13621688241239764

[31] Åsta Haukås. 2018. *Metacognition in Language Learning and Teaching*. Routledge, 11–30. doi:10.4324/9781351049146-2

[32] Sung-Hee Jin, Kowoon Im, Mina Yoo, Ido Roll, and Kyoungwon Seo. 2023. Supporting students' self-regulated learning in online learning using artificial intelligence applications. *International Journal of Educational Technology in Higher Education* 20, 1 (2023), 37.

[33] Binny Jose and Angel Thomas. 2025. Digital Anthropomorphism and the Psychology of Trust in Generative AI Tutors: An Opinion-Based Thematic Synthesis. *Frontiers in Computer Science* 7 (2025), 1638657. doi:10.3389/fcomp.2025.1638657

[34] Fatih Karataş, Faramarz Yaşar Abedi, Filiz Ozek Gunyel, Derya Karadeniz, and Yasemin Kuzgun. 2024. Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies* 29, 15 (2024), 19343–19366. doi:10.1007/s10639-024-12574-6

[35] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274. doi:10.1016/j.lindif.2023.102274

[36] Hassan Khosravi, Simon Buckingham Shum, Guanliang Chen, Cristina Conati, Dragan Gasevic, Judy Kay, Simon Knight, Roberto Martinez-Maldonado, Shazia Sadiq, and Yi-Shan Tsai. 2022. Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence* 3 (05 2022), 100074. doi:10.1016/j.caeai.2022.100074

[37] Eunhye Kim, Kiroong Choe, Minju Yoo, Sadat Shams Chowdhury, and Jinwook Seo. 2025. Beyond Tools: Understanding How Heavy Users Integrate LLMs into Everyday Tasks and Decision-Making. arXiv:2502.15395 [cs.HC] https://arxiv.org/abs/2502.15395

[38] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 822–835. doi:10.1145/3630106.3658941

[39] Malcolm Shepherd Knowles. 1975. *Self-directed learning*. Vol. 291. association press New York.

[40] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 54, 18 pages. doi:10.1145/3491102.3501999

[41] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (Nov. 2019), 35 pages. doi:10.1145/3359283

[42] Seongyong Lee, Hohsung Choe, Di Zou, and Jaeho Jeon. 2025. Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments* (2025), 1–25. doi:10.1080/10494820.2025.2498537

[43] Belle Li, Curtis J Bonk, and Xiaojing Kou. 2023. Exploring the multilingual applications of ChatGPT: Uncovering language learning affordances in YouTuber videos. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)* 13, 1 (2023), 1–22.

[44] Belle Li, Curtis J. Bonk, Chaoran Wang, and Xiaojing Kou. 2024. Reconceptualizing Self-Directed Learning in the Era of Generative AI: An Exploratory Analysis of Language Learning. *IEEE Transactions on Learning Technologies* 17 (2024), 1489–1503. doi:10.1109/TLT.2024.3386098

[45] Jingshu Li, Yitian Yang, Q. Vera Liao, Junti Zhang, and Yi-Chieh Lee. 2025. As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1111, 16 pages. doi:10.1145/3706598.3713336

[46] Tongguang Li, Debarshi Nath, Yixin Cheng, Yizhou Fan, Xinyu Li, Mladen Raković, Hassan Khosravi, Zachari Swiecki, Yi-Shan Tsai, and Dragan Gašević. 2025. Turning Real-Time Analytics into Adaptive Scaffolds for Self-Regulated Learning Using Generative Artificial Intelligence. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, 667–679. doi:10.1145/3706468.3706559

[47] Zixi Li, Chaoran Wang, and Curtis J Bonk. 2024. Exploring the Utility of ChatGPT for Self-Directed Online Language Learning. *Online Learning* 28, 3 (2024), 157–180. doi:10.24059/olj.v28i3.4497

[48] Lyn Lim, Maria Bannert, Joep van der Graaf, Shaveen Singh, Yizhou Fan, Surya Surendrannair, Mladen Rakovic, Inge Molenaar, Johanna Moore, and Dragan Gašević. 2023. Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior* 139 (2023), 107547. doi:10.1016/j.chb.2022.107547

[49] Xi Lin. 2024. Exploring the Role of ChatGPT as a Facilitator for Motivating Self-Directed Learning Among Adult Learners. *Adult Learning* 35, 3 (2024), 156–166. arXiv:https://doi.org/10.1177/10451595231184928 doi:10.1177/10451595231184928

[50] Huanyin Liu, Chengwei Lv, and Jianlin Chen. 2025. Willingness to communicate and anxiety in human–human and human–chatbot interaction contexts: an idiodynamic investigation. *Innovation in Language Learning and Teaching* 0, 0 (2025), 1–25. arXiv:https://doi.org/10.1080/17501229.2025.2560101 doi:10.1080/17501229.2025.2560101

[51] Meilu Liu, Lawrence Jun Zhang, and Christine Biebricher. 2024. Investigating students' cognitive processes in generative AI-assisted digital multimodal composing and traditional writing. *Computers & Education* 211 (2024), 104977.

[52] Sofie MM Loyens, Joshua Magda, and Remy MJP Rikers. 2008. Self-directed learning in problem-based learning and its relationships with self-regulated learning. *Educational psychology review* 20, 4 (2008), 411–427.

[53] Brian Lubars and Chenhao Tan. 2019. *Ask not what AI can do, but what AI should do: towards a framework of task delegability*. Curran Associates Inc., Red Hook, NY, USA.

[54] Santosh Mahapatra. 2024. Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments* 11, 1 (2024), 9. doi:10.1186/s40561-024-00295-9

[55] Rasulova Mahbuba. 2023. Unraveling the distinctions between self-directed learning and self-regulated learning. *International Journal of Advanced Multidisciplinary Research and Studies* 3 (2023), 1549–52.

[56] Sruti Mallik and Ahana Gangopadhyay. 2023. Proactive and reactive engagement of artificial intelligence methods for education: a review. *Frontiers in Artificial Intelligence* Volume 6 - 2023 (2023). doi:10.3389/frai.2023.1151391

[57] Memrise. 2022. Introducing MemBot, your new language partner! — memrise.com. https://www.memrise.com/blog/introducing-membot. [Accessed 01-12-2025].

[58] Sharan B Merriam and Lisa M Baumgartner. 2020. *Learning in adulthood: A comprehensive guide*. John Wiley & Sons.

[59] Anabil Munshi, Gautam Biswas, Ryan Baker, Jaclyn Ocumpaugh, Stephen Hutt, and Luc Paquette. 2023. Analysing adaptive scaffolds that help students develop self-regulated learning behaviours. *Journal of Computer Assisted Learning* 39, 2 (2023), 351–368. doi:10.1111/jcal.12761

[60] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 456, 20 pages. doi:10.1145/3706598.3713726

[61] Iris Cristina Peláez-Sánchez, Davis Velarde-Camaqui, and Leonardo David Glasserman-Morales. 2024. The impact of large language models on higher education: exploring the connection between AI and Education 4.0. *Frontiers in Education* Volume 9 - 2024 (2024). doi:10.3389/feduc.2024.1392091

[62] Nermin Punar Özçelik and Gonca Yangın Ekşi. 2024. Cultivating writing skills: the role of ChatGPT as a learning assistant—a case study. *Smart Learning Environments* 11, 1 (2024), 10. doi:10.1186/s40561-024-00296-8

[63] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 535, 14 pages. doi:10.1145/3491102.3501967

[64] r/languagelearning. 2025. r/languagelearning – Reddit community. https://www.reddit.com/r/languagelearning/. Accessed on 2025-06-09.

[65] Jennifer D. Robinson and Adam M. Persky. 2020. Developing Self-Directed Learners. *American Journal of Pharmaceutical Education* 84, 3 (March 2020), 847512. doi:10.5688/ajpe847512

[66] Kristina Schaaff and Marc-André Heidelmann. 2024. Impacts of Anthropomorphizing Large Language Models in Learning Environments. arXiv:2408.03945 [cs.CL] https://arxiv.org/abs/2408.03945

[67] Sarang Shaikh, Sule Yildirim Yayilgan, Blanka Klimova, and Marcel Pikhart. 2023. Assessing the usability of ChatGPT for formal English language learning. *European Journal of Investigation in Health, Psychology and Education* 13, 9 (2023), 1937–1960. doi:10.3390/ejihpe13090140

[68] Alexander M Sidorkin. 2025. AI integration blueprint: Transforming higher education for the age of intelligence. *AI-EDU Arxiv* (2025).

[69] Alexa Siu and Raymond Fok. 2025. Augmenting Expert Cognition in the Age of Generative AI: Insights from Document-Centric Knowledge Work. arXiv:2503.24334 [cs.HC] https://arxiv.org/abs/2503.24334

[70] Liyan Song and Janette R Hill. 2007. A Conceptual Model for Understanding Self-Directed Learning in Online Environments. *Journal of Interactive Online Learning* 6, 1 (2007), 27–42.

[71] Merrill Swain. 1985. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. *Input in second language acquisition* 15 (1985), 165–179.

[72] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. 2024. The Metacognitive Demands and Opportunities of Generative AI. In *CHI '24*. ACM. https://www.microsoft.com/en-us/research/publication/the-metacognitive-demands-and-opportunities-of-generative-ai/

[73] Wen-Ta Tseng, Zoltán Dörnyei, and Norbert Schmitt. 2006. A New Approach to Assessing Strategic Learning: The Case of Self-Regulation in Vocabulary Acquisition. *Applied Linguistics* 27, 1 (03 2006), 78–102. arXiv:https://academic.oup.com/applij/article-pdf/27/1/78/447478/ami046.pdf doi:10.1093/applin/ami046

[74] KURT VanLEHN. 2011. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist* 46, 4 (2011), 197–221. arXiv:https://doi.org/10.1080/00461520.2011.611369 doi:10.1080/00461520.2011.611369

[75] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (April 2023), 38 pages. doi:10.1145/3579605

[76] Virusnzz. 2025. Mod announcement: Lifting of the moratorium on AI apps. https://www.reddit.com/r/languagelearning/comments/1jcj7am/mod_announcement_lifting_of_the_moratorium_on_ai/. [Accessed 12-08-2025].

[77] Ben Wang. 2024. GOLF: Goal-Oriented Long-term liFe tasks supported by human-AI collaboration. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ACM, 3072–3072. doi:10.1145/3626772.3657655

[78] Chaoran Wang, Zixi Li, and Curtis Bonk. 2024. Understanding self-directed learning in AI-Assisted writing: A mixed methods study of postsecondary learners. *Computers and Education: Artificial Intelligence* 6 (2024), 100247. doi:10.1016/j.caeai.2024.100247

[79] Chenghao Wang, Bin Zou, Yiran Du, and Zixun Wang. 2024. The impact of different conversational generative AI chatbots on EFL learners: An analysis of willingness to communicate, foreign language speaking anxiety, and self-perceived communicative competence. *System* 127 (2024), 103533. doi:10.1016/j.system.2024.103533

[80] Feifei Wang, Xiaohua Zhou, Kangxin Li, Alan C. K. Cheung, and Ming Tian. 2025. The effects of artificial intelligence-based interactive scaffolding on secondary students' speaking performance, goal setting, self-evaluation, and motivation in informal digital learning of English. *Interactive Learning Environments* 33, 7 (2025), 4633–4652. arXiv:https://doi.org/10.1080/10494820.2025.2470319 doi:10.1080/10494820.2025.2470319

[81] Xiangzhi Eric Wang, Zackary P. T. Sin, Ye Jia, Daniel Archer, Wynonna H. Y. Fong, Qing Li, and Chen Li. 2025. Can You Move These Over There? An LLM-based VR Mover for Supporting Object Manipulation. arXiv:2502.02201 [cs.HC] https://arxiv.org/abs/2502.02201

[82] Yu Wang. 2025. Reducing anxiety, promoting enjoyment and enhancing overall English proficiency: The impact of AI-assisted language learning in Chinese EFL contexts. *British Educational Research Journal* (2025). doi:10.1002/berj.4187

[83] Mark Warschauer and Richard Kern. 2000. Introduction: Theory and practice of network-based language teaching. *Network-based language teaching: Concepts and practice* (2000), 1–19.

[84] Mark Warschauer, Waverly Tseng, Soobin Yim, Thomas Webster, Sharin Jacob, Qian Du, and Tamara Tate. 2023. The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Journal of Second Language Writing* 62 (2023).

[85] Anita L Wenden. 1998. Metacognitive knowledge and language learning. *Applied linguistics* 19, 4 (1998), 515–537.

[86] Watcharapol Wiboolyasarin, Kanokpan Wiboolyasarin, Kanpabhat Suwanwihok, Nattawut Jinowat, and Renu Muenjanchoey. 2024. Synergizing collaborative writing and AI feedback: An investigation into enhancing L2 writing proficiency in wiki-based environments. *Computers and Education: Artificial Intelligence* 6 (2024), 100228. doi:10.1016/j.caeai.2024.100228

[87] Da Yan. 2023. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies* 28, 11 (2023), 13943–13967. doi:10.1007/s10639-023-11742-4

[88] Lixiang Yan, Samuel Greiff, Ziwen Teuber, and Dragan Gašević. 2024. Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour* 8, 10 (2024), 1839–1850. doi:10.1038/s41562-024-02004-5

[89] Lu Yang and Rui Li. 2024. ChatGPT for L2 learning: Current status and implications. *System* 124 (2024), 103351. doi:10.1016/j.system.2024.103351

[90] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (Nov. 2018), 23 pages. doi:10.1145/3274463

[91] Barry J Zimmerman. 2000. Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation*. Elsevier, 13–39.

[92] Barry J Zimmerman and Magda Campillo. 2003. Motivating self-regulated problem solvers. *The psychology of problem solving* 233262 (2003), 103.