university of groningen
faculty of science and engineering
artificial intelligence

Architecture and Testing of Memory-Augmented Conversational Instructional Agents for Language Learning

Bachelor's Project Thesis

William Hermann, s5200954, w.hermann@student.rug.nl,
Supervisors: Paul A. Vogt & Suzan d. Scheffer

Abstract: Recent advances in conversational AI present new opportunities for language learning, yet current systems fail to integrate established second language acquisition (SLA) principles. This thesis addresses this gap by designing and implementing a multi-agent conversational AI system that combines authentic dialogue practice with evidence-based SLA methods, particularly spaced repetition. The system uses specialized agents to integrate SLA principles for vocabulary tracking, scheduled review, and authentic conversation. A pilot study using a modified experimental system compared this approach against traditional flashcard learning, assessing vocabulary recall and learner motivation. Results found comparable learning outcomes between conditions (conversational: 62.7%; flashcard: 70.9%; $p = .281$, $r = 0.357$) with small-to-moderate effect sizes for the conversational system in attention and satisfaction ($r \approx 0.40$) metrics. This work provides a working technical foundation for adaptive conversational learning systems and initial empirical insights into their educational potential.

1 Introduction

Research shows that incidental vocabulary acquisition through conversation can be 1.7 to 12 times more efficient per minute than direct instruction, indicating the potential of AI-powered conversational learning systems (McQuillan, 2019). This efficiency may come from social interaction, as research suggests meaning negotiation creates effective encoding conditions for long-term memory formation (Bitchener, 2004).

However, existing implementations fail to capture this potential because they lack the theoretical frameworks and goal-directed architectures needed to optimize vocabulary introduction, track learner progress, and implement evidence-based retention strategies (Belda-Medina & Calvo-Ferrer, 2022; Du & Daniel, 2024). While conversational practice provides the most natural environment for developing vocabulary knowledge, existing chatbots operate reactively rather than pedagogically. This prevents them from systematically targeting vocabulary or pursuing learning goals, creating an "adaptivity gap" between technical capability and educational effectiveness (Bibauw et al., 2022). Current systems cannot track individual learning progress or guide vocabulary toward specific goals.

Bridging this gap is important for two reasons. First, usage-based theories suggest that conversational systems offer unique learning advantages through social entrenchment. Schmid's Entrenchment and Conventionalization (EC)-Model proposes that linguistic knowledge is "continuously refreshed and reorganized under the influence of social interactions" (Schmid, 2017). From this perspective, entrenchment through communicative events may contribute to more robust memory formation than isolated study (Schmid, 2020). Second, evidence-based retention strategies could dramatically improve outcomes. Implementing spaced repetition algorithms can improve vocabulary recall by 29 percentage points (Belardi et al., 2021), yet this powerful technique remains untapped in conversational systems. To date, no peer-reviewed evaluations have been published of systems that integrate natural conversation with learner-specific vocabulary tracking and scheduled review, leaving these complementary advantages uncombined.

While entrenchment theory motivates the conversational approach, building effective vocabulary learning systems requires integrating multiple theoretical frameworks that address specific design challenges.

Theoretical foundations. CALL research shows that "coherent combinations of theories, or theory ensembles" are needed to accommodate the intersection of SLA, technology, and other disciplines (Hubbard & Levy, 2016). Four theories specifically address the design challenges of building pedagogically effective conversational systems, each informing key architectural decisions.

The New Theory of Disuse (Bjork & Bjork, 1992) reveals why neither conversation alone nor traditional methods alone is sufficient. The theory distinguishes storage strength (how well something is learned) from retrieval strength (how easily it can be recalled right now). When a word is learned, both increase, but retrieval strength fades quickly without practice while storage strength persists. This distinction exposes complementary limitations in existing approaches. Flashcard methods excel at building storage strength through systematic spacing, but they cannot develop retrieval pathways for spontaneous vocabulary use in authentic communication. Conversely, conversation develops these retrieval pathways through active use, but without systematic spacing it fails to optimize storage strength. An effective system must integrate both: spaced repetition algorithms to build storage strength, and conversational practice to develop retrieval pathways and procedural fluency necessary for spontaneous communication.

Skill Acquisition Theory (DeKeyser, 2007) explains why vocabulary learning requires different approaches at different stages. The theory describes progression from declarative knowledge (knowing meanings) through procedural knowledge (effortful use) to automatic fluency (unconscious application). This progression directly informs system architecture: newly introduced vocabulary needs explicit presentation with clear definitions and examples, while previously learned words require varied conversational practice to develop procedural fluency. Only extensive authentic speaking practice achieves the automatization necessary for real communication. This staged approach reflects Nation's distinction between vocabulary knowledge types, from form-meaning recognition to productive use (Nation, 2013).

The Interaction and Noticing hypotheses together inform authentic conversation design. The Interaction Hypothesis (Long, 1996) specifies that acquisition occurs through meaning negotiation in genuine communicative contexts, requiring authentic information gaps rather than scripted exchanges. Long's foundational work establishes that social interaction through meaning negotiation not only facilitates language acquisition but also creates superior encoding conditions for long-term memory formation. Research building on this theory has demonstrated that learners engaged in negotiation of meaning show significant advantages for both short-term and long-term memory compared to those in non-interactive conditions (Bitchener, 2004). The Noticing Hypothesis (Schmidt, 1990) complements this by establishing that conscious attention to vocabulary is necessary; words that pass unnoticed cannot be learned. Together, these theories justify systems that balance authentic conversation with deliberate vocabulary exposure, introducing words explicitly then using them in meaningful dialogue.

Together, these perspectives motivate a set of design considerations for vocabulary learning systems: tracking progression through skill stages (declarative to automatic), supporting conscious attention to target vocabulary, enabling authentic meaning negotiation in goal-oriented dialogue, and balancing storage and retrieval processes through strategic timing. The following section shows how these requirements can be addressed through system design.

Proposed approach. This study presents a multi-agent architecture designed to address the adaptivity gap by distributing the ensemble of theoretical requirements across specialized components, enabling natural conversation alongside systematic vocabulary tracking. To establish a baseline for measuring these architectural innovations, a preliminary experiment was conducted comparing the system against traditional flashcards. Flashcards serve as the established standard for declarative vocabulary testing, efficiently training word-meaning associations (Nation, 2013; Teymouri, 2024).

Research Question. How does vocabulary learning through multi-agent conversational practice compare to traditional flashcards in terms of immediate recall performance and learner motivation within a single-session experiment?

To answer this question, the system was developed in two versions:

1. A **primary version** using Neo4j (a graph database) knowledge graphs for vocabulary representation with spaced repetition scheduling based on forgetting curves, incorporating spacing intervals validated by Belardi et al. (2021).
2. An **experimental version** for controlled testing, using simplified CSV-based storage and priority scheduling to fit experimental time constraints.

The primary version requires weeks of testing to validate storage strength development. This study therefore uses an experimental adaptation that can test the core architectural elements within a single session: multi-agent coordination, systematic vocabulary selection, skill stage differentiation, and the attentional and motivational effects of conversational practice. While this single session cannot validate long-term memory entrenchment, it fits experimental constraints while maintaining the ability to evaluate the multi-agent approach.

The experiment compared the conversational system and flashcard practice for C2-level English vocabulary acquisition. Using a within-subjects design, both recall performance and intrinsic motivation (RIMMS scale) were measured and analyzed using non-parametric statistical tests to test for differences and generate hypotheses for future research (Loorbach et al., 2015).

Contributions of this study include:

1. **Technical implementation**: A working implementation of a multi-agent conversational vocabulary learning system that demonstrates how theoretical requirements can be addressed through distributed architecture.
2. **Preliminary empirical insights**: Initial evidence about performance patterns and motivational differences between conversational and traditional learning methods, establishing a methodological foundation for future controlled studies.

2 System Design

Building upon the theoretical requirements established in Section 1, this section presents the multi-agent architecture designed to address the adaptivity gap in conversational vocabulary learning. The section first describes the primary implementation designed for long-term deployment, then presents the experimental adaptation used in the pilot study (Section 2.6).

2.1 Architecture Overview

The system employs a multi-agent architecture comprising three specialized components: a conversation agent maintains natural dialogue flow, a language processor analyzes learner utterances, and a supervisor generates learning objectives. This specialization addresses performance degradation when single LLMs handle multiple complex tasks (Liu et al., 2024). Each component optimizes its primary function while maintaining coordinated system behavior. Detailed component descriptions follow in Section 2.3.

These components operate concurrently to preserve conversational flow, preparing learning goals and updating knowledge representations in the background. The conversation agent handles all user interaction, while the supervisor provides learning goals and the processor updates vocabulary knowledge in the background.

This division of responsibilities is especially important when implementing learning theories that create competing teaching demands.

2.2 Theoretical Connection

The four theoretical frameworks create implementation challenges that single-system architectures struggle to address simultaneously. Natural conversation flow (Interaction Hypothesis) can compete with explicit vocabulary attention (Noticing Hypothesis) for system focus; optimal spacing intervals (New Theory of Disuse) can interrupt con-

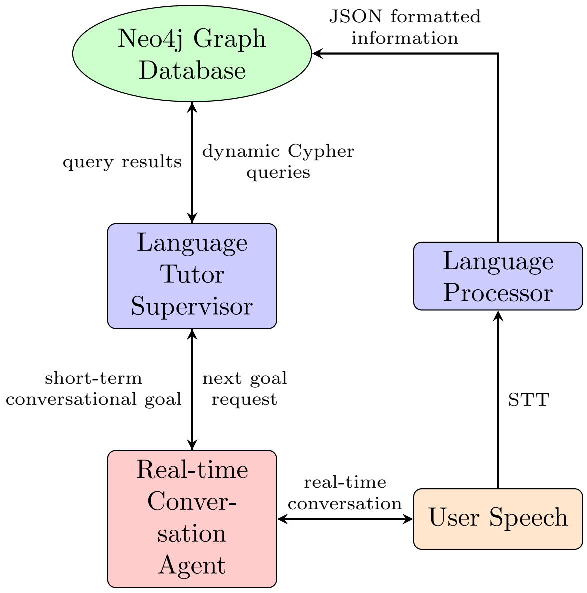


Figure 2.1: Multi-agent architecture with three core components operating concurrently. STT = speech-to-text.

versational practice needed for skill development (Skill Acquisition Theory). The multi-agent approach distributes these challenges across specialized components.

Each agent specializes while coordinating through shared knowledge: the conversation agent maintains dialogue flow and procedural practice; the supervisor orchestrates vocabulary selection, spacing, and noticing conditions; the language processor tracks skill development and performance indicators. The graph database integrates these observations, maintaining relationships between vocabulary, learning stages, and performance history. This allows each component to optimize its function without compromising others; conversation continues uninterrupted while learning decisions happen asynchronously.

Realizing this architecture requires technical infrastructure supporting real-time voice conversation and multi-agent coordination.

2.3 Core Components

Real-time Conversation Agent As the agent that users interact with directly, this component requires careful behavioral design to maintain authentic dialogue while providing procedural practice opportunities. The agent operates with extensive instructions that specify behaviors including error correction, goal-oriented conversation steering, and seamless integration with target vocabulary. These instructions emphasize sustained engagement across multiple sessions, balancing natural conversational flow with learning objectives. The complete instruction set is available in the source code repository (see Appendix A). The agent follows a goal-completion cycle: it receives learning objectives from the supervisor, integrates target vocabulary into conversation naturally, accomplishes these objectives, then requests new goals when ready. This maintains uninterrupted learning momentum while preserving conversational flow. Because complex reasoning is handled by the supervisor, this interface can operate effectively using a smaller, faster model, which also provides significant cost advantages. The Real-time Conversation Agent uses gpt-4o-mini-realtime through a WebSocket-based connection for ultra-low latency (1/20th the cost per minute of gpt-4o-realtime) with minimal performance loss. This separation reduces API costs from approximately \$50 to \$3 per hour of conversation while maintaining conversational quality, using the more powerful model only for intermittent planning tasks. OpenAI independently adopted this separation pattern in June 2025, validating the cost optimization approach.

Language Tutor Supervisor This supervisor addresses the adaptivity gap by tracking learner knowledge and adjusting content accordingly. It implements storage/retrieval optimization, grammatical pattern targeting, and skill progression management, generating comprehensive learning objectives that include vocabulary, grammar practice, and skill development based on individual progress. Operating with gpt-4o for superior reasoning, the supervisor processes recent conversation and database info without real-time constraints, activating only when new objectives are needed. This selective activation minimizes costs while preserving advanced decision-making for critical teaching moments.

When the conversation agent requests new goals, the supervisor analyzes context, generates tailored Cypher queries for the Neo4j database, retrieves relevant information (words due for review, grammatical patterns needing practice, appropriate new vocabulary), then synthesizes this data into actionable objectives like "naturally use: obfuscate, capitulate, pragmatic." The workflow introduces a one-second pause, providing intelligent direction without disrupting dialogue flow.

Language Processor Operating as a silent overseer, this processor maintains database currency by piggybacking on the user-side conversation transcript from the Real-time Conversation Agent and transforming it into structured learning data. It analyzes all user utterances, identifying grammatical features and correctness for each word, then outputs findings in database-compatible JSON schema. This passive observation enables comprehensive linguistic analysis without disrupting conversation; the processor influences learning only indirectly through database updates that inform future supervisor decisions.

2.4 Knowledge Graph and Memory System

Graph Structure Neo4j was selected to model vocabulary learning's complex relational structure: words connect to multiple grammatical forms, each with distinct performance histories and spacing schedules. Graph databases excel at this interconnected data, where queries like "words where user struggles with genitive case" require simple traversals but would need complex SQL Joins in traditional databases.

The structure models vocabulary learning through interconnected nodes: `Lexeme` nodes represent vocabulary items, connecting to `GrammarContext` nodes via `USES_GRAMMAR` relationships. Grammar contexts link to morphological features (`HAS_CASE`, `HAS_TENSE`, `HAS_NUMBER`), enabling form-specific performance tracking, distinguishing, for example, nominative mastery from genitive difficulty within the same word.

The core learning data resides in `LearningProgress` nodes, which connect users to vocabulary items through the relationship path: User-HAS_PROGRESS-LearningProgress-ABOUT-Lexeme. These nodes unify spaced repetition scheduling with performance analytics.

Learning Progress Integration Each `LearningProgress` node contains:

- **Scheduling data**: SRS level (1-5 Leitner box system), next review date, last interaction timestamp
- **Performance metrics**: Overall success rates, form-specific statistics (JSON), weakest forms array, cumulative encounter/success counts

The Leitner box system was selected over advanced algorithms like SM-2 and half-life regression due to conversational constraints. These algorithms require detailed self-reported difficulty ratings to achieve their claimed accuracy. Initial attempts to estimate these ratings by having an LLM analyze conversational interactions proved inaccurate, while explicitly asking users for ratings would interrupt conversation flow. This left binary success/failure classification as the only viable approach. The Leitner box system is well-suited to this constraint and validated by studies showing significant vocabulary gains (Davatgar & Ghorbanzadeh, 2013), providing adequate functionality for architecture validation.

The adaptive mechanism uses exponential spacing intervals (1, 2, 4, 8, 16 days) that expand as proficiency increases: struggling words return to shorter intervals through SRS reduction, while mastered items advance to longer review periods. This system operationalizes the New Theory of Disuse (Bjork & Bjork, 1992) by distinguishing storage strength from retrieval strength. Storage strength builds through cumulative encounters and SRS progression, while retrieval strength is enhanced through the Neo4j graph structure that enables traversal across related grammatical forms and contextual connections. The system classifies each interaction as accurate usage, morphological error, or retrieval failure, then continuously updates both strength metrics to inform scheduling decisions.

2.5 Technical Foundation

The system builds on OpenAI's Swarm framework for its real-time voice capabilities, which provide pre-built WebSocket infrastructure for audio streaming, interruption detection, and turn management. Building this infrastructure from scratch would require complex state synchronization and audio buffer handling, making Swarm the optimal choice for rapid development.

Swarm's function calling capabilities enable agents to execute predefined code functions during conversation without breaking dialogue flow. In this architecture, function calls serve as the primary mechanism for the conversation agent to request new learning goals from external systems. The framework also provides agent handoff capabilities that enable different agents to control the conversation at different times, a feature utilized in the experimental adaptation (Section 2.6).

However, the primary architecture required parallel processing capabilities that Swarm was not designed to support. The framework assumes sequential control: one agent handles the conversation, then passes complete control to another. To implement parallel operation, the system uses shared context windows and context routing between agents, with separate processing threads enabling asynchronous processing. This allows the conversation agent to maintain real-time dialogue while the supervisor and processor work in parallel, accessing shared conversation context as needed. Modifying Swarm's backend to support simultaneous operation of three agents working on the same conversation stream was challenging but necessary, representing the optimal trade-off between development speed and architectural requirements.

2.6 Experimental Adaptation

The preceding sections described the primary implementation designed for long-term deployment with weeks of interaction. However, as noted in Section 1, this study also developed an experimental version to enable controlled single-session testing. While the primary version demonstrates full theoretical implementation, controlled experimental validation required architectural adaptation to fit single-session experiment constraints. Preserving multi-agent coordination and theoretical principles within experimental logistics (35-minute sessions, pretest integration, block-based design, and avoiding Neo4j database setup for each participant) required simplified but functionally equivalent implementations.

Two-Agent Handoff System To better target skill acquisition stages within the short experimental window, the Real-time Conversation Agent is replaced by two specialized voice agents that make use of OpenAI Swarm's handoff capability (Figure 2.2). This approach directly implements DeKeyser's Skill Acquisition Theory (DeKeyser, 2007), dividing vocabulary learning between declarative and procedural phases for more precise targeting within the 35-minute constraints.

The IntroducerAgent specializes in the declarative stage, presenting definitions and contextual examples to establish explicit knowledge. The ReviewAgent focuses on procedural development and entrenchment, practicing words in varied conversational contexts. Complete instruction sets for both agents are available in the source code repository (see Appendix A).

This architectural separation prevents instruction dilution across DeKeyser stages and provides natural learning queue management. Similar to how flashcard users control their acquisition pace, the two-agent system creates a more structured approach to vocabulary introduction and review timing.

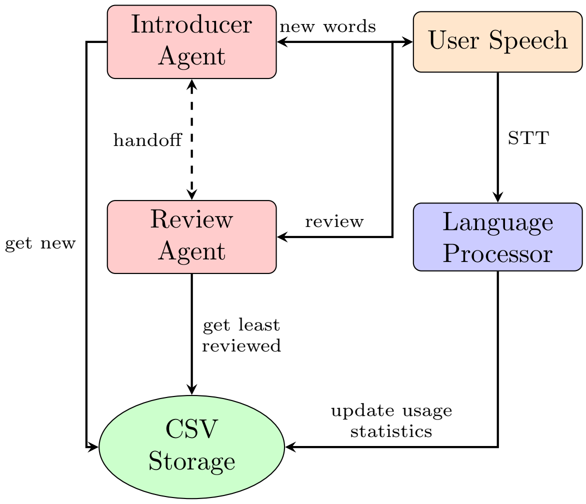


Figure 2.2: Experimental two-agent handoff system implementing DeKeyser's skill acquisition stages. Introducer and Review agents are specialized versions of core components (see Section 2.3). Verbal commands trigger transitions between modes.

Interaction Design While the primary architecture emphasizes sustained engagement for long-term use, the experimental timeframe required intensifying these engagement principles into a high-throughput "personal trainer" style: maximum speed, energetic delivery, and focused interaction. This intensification addresses throughput constraints: with 10 words to learn across two 6-minute blocks (12 minutes total per condition), the system required exceptional learning efficiency to compete with flashcard methods on declarative assessments. Flashcards purely train declarative memory without conversational overhead, making timing a critical constraint for conversational approaches that naturally introduce inefficiencies. The focused, directive interaction style prevents time loss to unnecessary conversation, maintaining the pace necessary to achieve comparable declarative learning outcomes.

The decision to place handoff control with participants addressed experimental constraints: with 6-minute blocks and no learner models, participants could better assess their cognitive load than an automated system. Learners progressed at different speeds: some needed more declarative introduction, others benefited from rapid procedural practice. The verbal commands "I want new words" or "I want to review words" triggered handoff, allowing self-regulation and preventing frustration from being overwhelmed or reviewing mastered words. This design was crucial since without weeks of interaction data, the system relied on participants' metacognitive awareness for optimization within brief sessions. The agent separation ensured focused declarative learning while review efficiently practiced multiple words for procedural development.

The user-controlled design also offered motivational benefits. According to Self-Determination Theory, autonomy is one of three core psychological needs that drive human motivation (Ryan & Deci, 2000). When learners control when to introduce new words versus practice familiar ones, they feel more invested in the process. Instead of simply following a prescribed sequence, participants actively shape their learning path. This sense of collaboration and ownership may enhance engagement, as learners feel they are working with the system rather than being directed by it.

Simplified Coordination Instead of Neo4j graph databases and complex spacing algorithms, the experimental system used CSV files with priority-based scheduling. The Language Tutor Supervisor's complex database queries were replaced by simple tool calls that execute Python scripts to pull words from priority queues. When agents requested new words or review words, getNewWords or getReviewWords tool calls ran scripts selecting vocabulary based on priority: recently introduced words with lowest user frequency received priority for review, while unseen words were prioritized for introduction. The Language Processor continued analyzing conversations but generated CSV updates tracking usage statistics. This simplified approach eliminated the need for complex Cypher queries while maintaining systematic vocabulary selection within the 6-minute blocks, acknowledging that spaced repetition optimally requires weeks to demonstrate effectiveness.

3 Methods

3.1 Overview

This study employed a within-subjects design comparing memory-augmented conversational agents with traditional flashcard learning for vocabulary acquisition. Word recall performance and learner motivation were assessed using the reduced Instructional Materials Motivation Survey (RIMMS) (Loorbach et al., 2015). The experimental adaptation (Section 2.6) enabled single-session evaluation of the system.

3.2 Experimental Design

We employed a counterbalanced within-subjects design with two conditions:

- **Condition A (CONV):** Interaction with the conversational agent system using the experimental implementation, includes live transcription of conversation on display
- **Condition B (Control):** Traditional flashcard learning within a custom WebUI with definitions and 1-2 contextual examples per word. Participants can choose next word or to restart from the start of the list, so users can actively

choose to review words or learn new ones. (see Appendix C for flashcard interface)

Participants learned 20 words (10 per condition) across four 6-minute blocks with 30-second breaks. Counterbalancing used ABAB/BABA order based on participant number (odd: ABAB; even: BABA). The 6-minute duration enabled four blocks plus assessments within 35 minutes, balancing exposure with practical constraints. RIMMS assessments followed blocks three and four to optimize memory spacing and minimize survey fatigue. The ABAB structure introduced 6-8 minute forgetting intervals between exposures to each condition, allowing assessment of relearning after brief delays.

3.3 Participants

Twelve university students (18-26 years, C1-C2 English proficiency, non-native speakers) were recruited, with one excluded for incomplete posttest data, leaving eleven participants (7 male, 4 female; 5 ABAB, 6 BABA order).

3.4 Materials

Vocabulary Selection Twenty-nine C2-level English words plus one control word ("happy") formed the experimental corpus (see Appendix B for complete list).

Pretest Tool A GUI pretest presented all 30 vocabulary items (29 C2-level words plus the familiar word "happy"). Participants selected words they already knew, which were removed from their individual learning sets. The inclusion of "happy" served as an attention check; participants were required to select this known word to ensure they were reading instructions carefully. After the pretest, the Experimental Master Controller automatically selected 20 unknown words from each participant's remaining vocabulary and randomly assigned 10 to each experimental condition.

Experimental Master Controller A custom software tool was developed to automate and standardize the entire experimental procedure, ensuring identical experimental conditions for all participants. The software handled all aspects of the experiment: administering the vocabulary pretest, automatically generating personalized 20-word learning sets based on pretest results, loading the correct word lists into each learning system at the appropriate blocks, enforcing precise timing for all 6-minute learning blocks and 30-second breaks, triggering RIMMS surveys at predetermined points, and generating customized post-tests that matched each participant's assigned vocabulary.

3.5 Measures

Vocabulary Assessment A multiple-choice cloze posttest administered 12-24 hours post-session via Google Forms tested all 20 learned words. Questions assessed contextual understanding, for example: "The politician tried to ____ the facts to avoid taking responsibility for the scandal" with "obfuscate" among the options. This format assessed declarative vocabulary knowledge through contextualized recognition rather than productive use, following validated assessment procedures established by Read (Read, 2000). The assessment originally included productive use tasks, but pilot testing showed this was too lengthy for completion; the first participant's data was excluded and the assessment was simplified to multiple-choice only. This declarative-only format favors flashcard methods, which train declarative memory without conversational overhead, but provided a controlled comparison point for evaluating the conversational system's declarative learning efficiency. The assessments were automatically created from a question bank of 29 questions to match each participant's personalized vocabulary set.

Motivation Assessment The 12-item reduced Instructional Materials Motivation Survey (RIMMS) was used, derived from the full IMMS. It measured four dimensions (Attention, Relevance, Confidence, and Satisfaction) using 5-point Likert scales (Loorbach et al., 2015). Each dimension was assessed with targeted questions administered immediately after participants completed their second exposure to each condition.

3.6 Procedure

The experiment was conducted in-person in a quiet environment. Upon arrival, participants were seated at a computer workstation with the experimental software preloaded. The experimenter remained present throughout the session to ensure protocol adherence and address technical issues, but did not interact with participants during learning blocks.

After obtaining written informed consent, the experimenter provided standardized instructions explaining the experimental structure and both learning systems. For the conversational agent condition, participants were specifically instructed to control their learning by verbally requesting either "I want new words" or "I want to review words" to switch between vocabulary introduction and practice modes. Participants were encouraged to actively direct their learning pace and strategy using these commands.

The experimental session followed the structure shown in Figure 3.1. Participants first completed the vocabulary familiarity pretest, identifying known words from the C2 corpus. The Experimental Master Controller then generated their personalized 20-word learning set and assigned words to conditions.

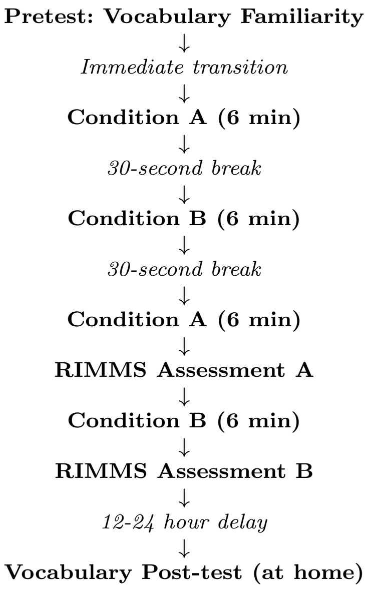


Figure 3.1: Complete experimental flow for ABAB condition order. BABA participants experienced the same structure with conditions reversed. Total session duration: approximately 35 minutes.

Learning blocks proceeded with automated transitions and timing enforced by the Master Controller. Visual and auditory signals indicated block transitions. During CONV blocks, participants engaged with the conversational agent through the web interface, while Control blocks presented flashcards with navigation controls.

After completing all four learning blocks and both RIMMS assessments, the experimenter debriefed participants, reminded them to take the posttest the next day, and answered any questions.

The vocabulary posttest was automatically sent via email 12 hours after session completion. Participants completed this assessment at home without supervision, using their personal devices. They had a 12-hour window to complete the assessment.

3.7 Data Analysis

Performance Analysis Learning effectiveness was assessed through delayed recall accuracy on the multiple-choice posttest. Due to small sample size (n=11), discrete performance data (5% increments resulting from 10 questions per condition), and non-normal distributions confirmed by Shapiro-Wilk tests, we employed Wilcoxon signed-rank tests to compare conditions. Effect sizes were calculated using r = $z/\sqrt{n}$ to quantify the magnitude of differences.

Motivation Analysis RIMMS responses were analyzed using Wilcoxon signed-rank tests for dimension-level comparisons between conditions. Additionally, correlation analysis examined relationships between motivational dimensions and performance outcomes to identify potential mediating factors.

Statistical Power Post-hoc power analysis revealed 34% power for the observed performance difference (r = 0.31) and 54% power for motivational effects (r ≈ 0.40). Approximately 21 participants would be required to achieve 80% power for detecting moderate effects (r = 0.50). Assuming normal distribution and using parametric tests with larger samples, approximately 74 participants would be needed for 80% power to detect the performance effect (r = 0.31), and 42 participants for the motivational effects (r = 0.40). The study was underpowered to detect small-to-moderate effects.

4 Results

This section presents experimental findings from eleven participants who completed the entire training phase and posttest.

4.1 Participant Characteristics

Eleven participants completed all experimental tasks, including vocabulary pretest, both learning sessions in counterbalanced order, RIMMS motivation assessments, and the 12-24-hour delayed posttest. Five participants received the ABAB condition order and six received the BABA order. All participants were C1-C2 level English speakers who knew fewer than 9 of the 29 target vocabulary words during pretesting, ensuring sufficient learning opportunity.

4.2 Learning Performance Outcomes

Overall Performance Comparison The 12-24 hour delayed vocabulary assessment revealed comparable retention patterns between learning conditions. Conversational learning achieved a mean accuracy of 62.7% (SD = 21.4%, Range: 20%-90%), while flashcard learning yielded 70.9% (SD = 17.3%, Range: 30%-100%).

We analyzed the data using non-parametric statistics due to the small sample size (n=11), discrete performance measurements (5% increments), and tied values. The Wilcoxon signed-rank test showed no significant difference between conditions (W = 7.0, p = .281, r = 0.357). This effect size should be interpreted with caution given non-significance and low statistical power. These results indicate comparable learning outcomes between methods within this pilot study's constraints (Figure 4.1).

Individual Learning Patterns Aggregate performance was similar, but individual responses revealed variations in performance patterns between conditions. Performance differences for individual participants ranged from -50% to +30%, indicating different learner responses to each method.

Three participants (P003, P006, P007) achieved 20-30% higher scores with the conversational agent, while four participants (P004, P008, P009, P010) performed 30-50% better with flashcards, with P010 achieving perfect recall. The remaining four participants (P002, P005, P011, P012) demonstrated similar performance in the 70-80% range regardless of method. This individual variability suggests learners may respond differently to each approach, though the small sample size prevents claims about stable learner characteristics. Without larger samples, these individual differences cannot be distinguished from random variation, though the systematic patterns (conversational advantages for some learners, flashcard advantages for others) warrant investigation in adequately powered future studies (Figure 4.2).

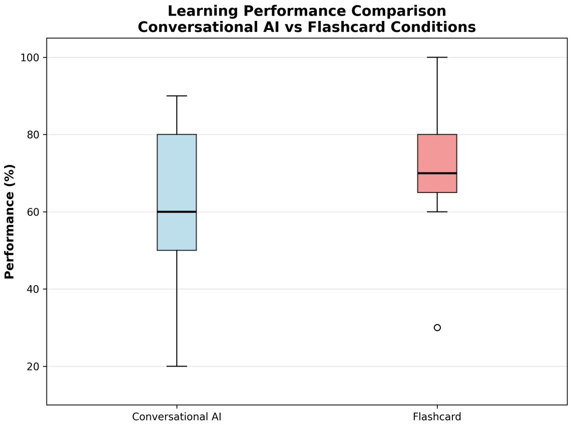


Figure 4.1: Learning performance comparison between conversational AI and flashcard conditions. Statistical analysis revealed no significant difference between conditions (Wilcoxon W = 7.0, p = .281, r = 0.357).

4.3 Motivational Response Analysis (RIMMS)

The Reduced Instructional Materials Motivation Survey showed differential effects across the four motivational dimensions. Statistical significance was not achieved due to sample size constraints. While effect size estimates were moderate (r ≈ 0.40) for attention and satisfaction dimensions, these should be interpreted as preliminary patterns rather than reliable effects given the non-significant results and limited statistical power.

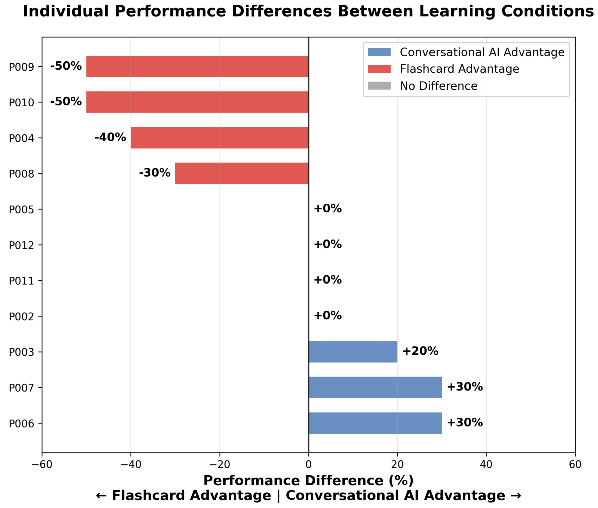


Figure 4.2: Individual performance differences between learning conditions. Horizontal bars extend from center showing the magnitude and direction of performance differences. Red bars indicate flashcard advantages, blue bars show conversational AI advantages. Participants are sorted by difference magnitude, revealing distinct individual learning preferences ranging from -50% to +30%.

Attention and Engagement Conversational agents had higher attention scores, with a mean of 3.88 (SD = 1.15) compared to 2.85 (SD = 1.13) for flashcards. Although the Wilcoxon signed-rank test did not reach significance (W = 17.5, p = .183), the moderate effect size estimate (r = 0.402) warrants investigation in adequately powered studies.

Satisfaction and User Experience Conversational learning produced higher satisfaction ratings (M = 3.70, SD = 1.00) than flashcard learning (M = 2.91, SD = 1.15). The difference was not statistically significant (W = 14.0, p = .188), but the moderate effect size estimate (r = 0.405) warrants further investigation.

Relevance and Confidence Both methods were perceived as equally relevant to learning goals (Conversational: M = 4.03, SD = 0.52; Flashcard: M = 3.94, SD = 0.55; W = 17.0, p = .961, r = 0.015). Similarly, learner confidence showed minimal differences between conditions (Conversational: M = 3.58, SD = 0.88; Flashcard: M = 3.33, SD = 0.57; W = 13.0, p = .539, r = 0.168). The non-significant patterns showed higher attention and satisfaction for conversational learning without corresponding gains in relevance or confidence (Figure 4.3).

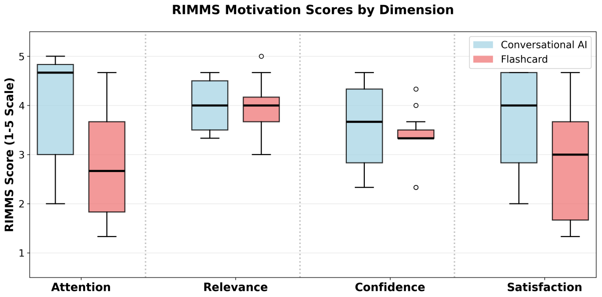


Figure 4.3: RIMMS motivation score distributions across four dimensions comparing conversational AI and flashcard conditions. Box plots show medians, quartiles, and outliers for each condition. Conversational AI showed higher attention and satisfaction scores, while relevance and confidence showed minimal differences.

5 Discussion

5.1 Key Findings

This pilot study found no significant performance differences between conversational and flashcard learning (p = .281), though comparable mean scores (conversational: 62.7%; flashcard: 70.9%) demonstrate the system functioned as designed. The conversation agent, supervisor, and processor coordinated successfully within the experimental constraints, establishing technical feasibility. This architecture may close the adaptivity gap, but adequately powered studies are needed to test its intended conditions: extended deployment, spaced repetition, procedural skill development, and retrieval strength assessment.

The RIMMS analysis showed no significant differences, though moderate effect size estimates (r ≈ 0.40) for attention and satisfaction warrant investigation in larger samples. Higher attention and satisfaction without confidence gains aligns with theory suggesting that engaging interaction requires extended exposure before perceived mastery develops, though this remains speculative given non-significant results.

5.2 Methodological Constraints and Their Implications

This study's outcomes were limited by several methodological constraints that shape interpretation of the findings.

Statistical Power The most significant limitation was the small sample size (n=11), which provided insufficient statistical power to detect differences between conditions. This constraint makes it difficult to distinguish genuine patterns from individual variation and limits generalizability.

Training Asymmetry Participants often brought years of flashcard experience but received no prior training for AI conversational learning. Transfer learning from human conversation skills may have benefited the conversational condition, but some participants reported intimidation with the AI system, potentially offsetting these natural advantages. This experience gap may have favored familiar methods over novel approaches.

Single-Session Experimental Constraints The most severe limitation is conducting a single-session experiment for an architecture designed for multi-week deployment. The experimental version used simplified components for 35-minute sessions, but the full theoretical advantages require weeks of interaction to emerge. The 12-minute exposure per condition is insufficient for vocabulary acquisition research, as memory consolidation, spaced repetition benefits, and vocabulary entrenchment require repeated exposure over extended periods. This limited time frame tests only immediate familiarity rather than genuine learning and prevents the emergence of conversational learning's theoretical advantages through social interaction and systematic spacing. The ABAB design introduced forgetting intervals, but they were insufficient for substantial forgetting or proper evaluation of long-term retention.

Assessment Limitations The assessment approach primarily tested storage strength rather than retrieval strength. The New Theory of Disuse (Bjork & Bjork, 1992) explains why flashcard methods excel on declarative recognition tests: they systematically build storage strength through repeated exposure and strategic spacing, which multiple-choice assessments directly measure. Conversational practice develops retrieval pathways for spontaneous vocabulary use in authentic communication, which require production tasks to properly evaluate.

Within the tight 6-minute blocks, the two conditions used time differently. Flashcards dedicated full time to vocabulary exposure. Conversational learning split time between vocabulary and conversation mechanics (turn-taking, dialogue flow, social interaction), resulting in less pure vocabulary exposure per word. As described in Section 2.6, the experimental design used focused, directive agent instructions to minimize this overhead and maintain learning efficiency. Despite these efforts to optimize throughput, the time split remained. The multiple-choice assessment tested only declarative knowledge, excluding both the retrieval strength and procedural fluency that conversation develops. Testing primarily through declarative recognition while conversation received less vocabulary exposure time fundamentally limited evaluation of conversational learning's theoretical advantages.

References

Belardi, A., Pedrett, S., Rothen, N., & Reber, T. P. (2021). Spacing, feedback, and testing boost vocabulary learning in a web application. *Frontiers in Psychology*, *12*, 757262.

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as ai conversational partners in language learning. *Applied Sciences*, *12*(17), 8427. doi: 10.3390/app12178427

Bibauw, S., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, *26*(1), 1–24.

Bitchener, J. (2004). The relationship between the negotiation of meaning and language learning: A longitudinal study. *Language Awareness*, *13*(2), 81–95.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In *From learning processes to cognitive processes: Essays in honor of william k. estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davatgar, H., & Ghorbanzadeh, Z. (2013). The effect of leitner's learning box on the improvement of vocabulary teaching and learning (case study: First year students in parsabad moghan branch, islamic azad university, parsabad moghan, iran). *IBU Repository*.

DeKeyser, R. M. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 97–113). Mahwah, NJ: Lawrence Erlbaum Associates.

Du, J., & Daniel, B. K. (2024). Transforming language education: A systematic review of ai-powered chatbots for english as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, *6*, 100230. doi: 10.1016/j.caeai.2024.100230

Hubbard, P., & Levy, M. (2016). Theory in computer-assisted language learning research and practice. In *The routledge handbook of language learning and technology* (pp. 24–38). Routledge.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. Retrieved from `https://aclanthology.org/2024.tacl-1.9/` doi: 10.1162/tacl_a_00638

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (Vol. 2, pp. 413–468). New York: Academic Press.

Loorbach, N., Peters, O., Karreman, J., & Steehouder, M. F. (2015). Validation of the instructional materials motivation survey (imms) in a self-directed instructional setting aimed at working with technology. *British Journal of Educational Technology*, *46*(1), 204–218. doi: 10.1111/bjet.12138

McQuillan, J. L. (2019). The inefficiency of vocabulary instruction. *International Electronic Journal of Elementary Education*, *11*(4), 309–318.

Nation, I. (2013). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78.

Schmid, H.-J. (Ed.). (2017). *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge.* Malden, MA and Berlin: American Psychological Association and De Gruyter.

Schmid, H.-J. (2020). *The dynamics of the linguistic system: Usage, conventionalization, and entrenchment.* Oxford: Oxford University Press.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158.

Teymouri, R. (2024). Recent developments in mobile-assisted vocabulary learning: A mini review of published studies focusing on digital flashcards. *Frontiers in Education*, *9*(1496578), 1–14. doi: 10.3389/feduc.2024.1496578

A Code Repositories

The complete source code for this research is available in the following GitHub repositories:

Primary System Implementation: `https://github.com/coding-crying/realtime-agents-language-tutor/tree/main`
Neo4j graph database implementation with spaced repetition algorithms and real-time voice agent integration.

Experimental Version & Data Analysis: `https://github.com/coding-crying/S5200954-Bachelor-Project-/tree/main`
Simplified CSV-based implementation used in the experimental evaluation, including data analysis scripts, statistical computations, and visualization code.

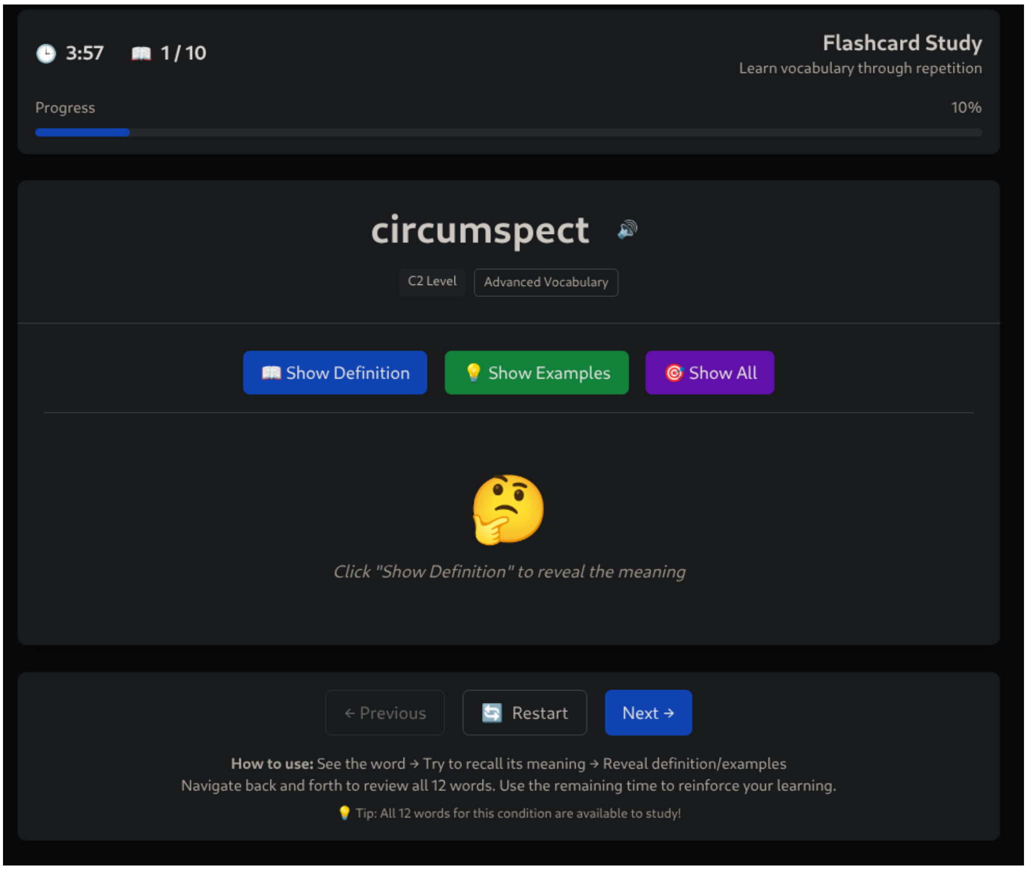
Agent instructions for the experimental system: `https://github.com/coding-crying/S5200954-Bachelor-Project-/blob/main/src/app/agentConfigs/vocabularyInstructor/index.ts`

B Experimental Vocabulary List

obfuscate, disparage, perfunctory, precocious, quandary, circumspect, capitulate, vociferous, intractable, abrogate, abstruse, acumen, admonish, austere, bolster, cacophony, cajole, candor, capricious, conciliatory, conundrum, copious, cursory, deleterious, ephemeral, eschew, garrulous, hackneyed

Control word: happy

C Flashcard Learning Interface



Click "Show Definition" to reveal the meaning

How to use: See the word → Try to recall its meaning → Reveal definition/examples
Navigate back and forth to review all 12 words. Use the remaining time to reinforce your learning.

💡 Tip: All 12 words for this condition are available to study!

Figure C.1: Flashcard interface showing vocabulary word with definition and contextual examples