1RT730 Project Report

Harald - Your Personal Chat-Powered Language Partner

Joel Sundin, Petter Möllerström, Rahul Sebastian Peter

October 20, 2025

1 Introduction

Learning Swedish as a non-native speaker presents significant challenges, particularly when opportunities for interactive practice and real-time feedback are limited. To address these challenges, this work proposes *Harald*, an AI-driven language learning application that leverages Large Language Models (LLMs), quizzes, and flashcard functionalities to enhance engagement and personalization in language acquisition.

Harald integrates multiple learning modalities to support comprehension, vocabulary retention, and conversational skills. Beyond individual learning, Harald contributes to societal language inclusivity by supporting immigrants, students, and professionals in their integration into Swedish society. The primary contribution of this study lies in the integration of LLM-based interaction within a structured learning framework, enabling personalized practice, targeted feedback, and contextual understanding in a scalable and accessible manner.

Harald is implemented as a web-based application, allowing learners to access its features directly through a standard browser without the need for local installation. This design choice enhances accessibility and convenience, enabling learners to practice Swedish anytime and anywhere, whether on desktop or mobile devices. By combining cloud-based LLM computation with an intuitive browser interface, Harald ensures that advanced language processing capabilities are delivered seamlessly to users, while maintaining a lightweight and user-friendly experience.

The system's web-based nature also facilitates easy updates and integration of new exercises or models, ensuring that learners always benefit from the latest functionality Harald has to offer. Overall, Harald provides a comprehensive, adaptive, and accessible environment for Swedish language learning, bridging the gap between traditional pedagogical methods and modern AI-assisted instruction.

2 Dataset

The development of the Swedish language tutoring chatbot involved several iterations of dataset design and experimentation aimed at improving the model's performance, naturalness, and pedagogical value. Over the course of the project, multiple datasets were constructed and evaluated, ranging from community-sourced linguistic discussions to synthetically generated dialogue data. Ultimately, however, the final system adopted a different strategy that relied on prompt engineering and curated conversational examples rather than direct fine-tuning. The main dataset iterations are described below.

2.1 Reddit-based corpus

The initial dataset was collected from the r/swedish subreddit using the Python Reddit API Wrapper (PRAW). This community consists of users interested in learning or discussing the Swedish language, and thus provides an organic dataset composed of learner questions and native or advanced speaker responses. The dataset comprised approximately 6,500 comments, each representing short-form discussions about grammar, vocabulary, pronunciation, and translation. Preliminary fine-tuning experiments were conducted using this dataset to adapt a general-purpose language model to the linguistic style and content of Swedish language learners. However, despite its relevance, the dataset exhibited several limitations, including inconsistent quality, informal tone, and limited conversational depth. Consequently, it was deemed insufficient as a primary fine-tuning source.

2.2 Synthetic dialogue dataset

In the second iteration, a manually constructed dataset of user–assistant interactions was developed to simulate tutoring dialogues. These pairs were designed to represent common learner scenarios such as vocabulary practice, grammar correction, and cultural explanations. This dataset was later expanded using generative augmentation, where additional examples were synthesized with Anthropic's Claude

Sonnet model. The resulting dataset provided higher-quality, domain-specific examples and was used for initial LoRA-based fine-tuning experiments. While the synthetic data improved linguistic consistency, further evaluation revealed limited benefits relative to the computational cost of model adaptation.

2.3 Web-scraped Swedish news dataset

To provide learners with authentic, up-to-date material for reading comprehension and writing practice, a web scraper was implemented to collect recent Swedish news articles from online media outlets. The system allows users to choose from a selection of articles and engage in follow-up exercises such as summarization or open-ended question answering. This component operates dynamically at runtime rather than as part of the fine-tuning dataset, ensuring continuous exposure to contemporary language usage and diverse topics.

2.4 Flashcard-based vocabulary dataset

An auxiliary feature enables users to store specific words encountered during interaction into a personal flashcard list for later review. Although not a dataset in the traditional training sense, this mechanism contributes to personalized learning and vocabulary retention through spaced repetition.

2.5 Final data strategy

In the final system, neither the Reddit corpus nor the synthetic fine-tuning datasets were employed in the deployed model. Instead, the chatbot is based on a fully pretrained large language model (Gemini 2.5 Flash), with its pedagogical behavior governed by an extensive system prompt. This prompt defines the tutor's persona, instructional goals, and tone, supplemented by curated example dialogues that demonstrate effective interaction patterns. This approach provided greater flexibility, reduced training costs, and preserved the model's general linguistic competence while enabling targeted guidance for Swedish language tutoring.

3 Architecture

The application is implemented in `Streamlit` following a modular, multi-page architecture, where each module corresponds to a specific learning activity. At its core, the application´s LLM capabilities is based on Google's *Gemini 2.5 Flash* model, while `Streamlit's session state` provides a lightweight mechanism for preserving user progress and contextual information across modules.

A minimal landing page serves as the system's entry point, directing users to one of four main environments: *Chat*, *Quiz*, *Comprehension*, and *Flashcards*. Although these pages operate independently, they share a common state management layer that allows data to persist seamlessly between activities.

The *Chat page* constitutes the central interface for interaction with Harald (LLM). Upon initialization, a composite system prompt is constructed that merges Harald's persona, sample dialogues, and contextual information derived from the learner's previous activity. When quiz data are available, the prompt automatically incorporates relevant performance metrics such as scores and incorrectly recalled vocabulary. All messages in the conversation is stored both in session memory and in an external JSON file to enable continuity across sessions. Communication with Gemini is handled through a `genai.Client` configured via `GenerateContentConfig` objects specifying prompt templates and temperature parameters.

The *Quiz page* delivers randomized multiple-choice assessments from a predefined dataset of questions. User responses are evaluated, summarized, and stored in the shared session state, thereby enriching the context used by Harald in subsequent interactions. This continuous exchange between tests and

conversation forms an adaptive feedback loop that connects assessment results with targeted language reinforcement.

The *Comprehension page* focuses on reading and writing proficiency by presenting short Swedish articles and supporting two modes of interaction: summarization and comprehension. In the summarization mode, learners write summaries that the model evaluates for grammatical accuracy and stylistic appropriateness. In the comprehension mode, the model generates questions about the article, records the learner's answers, and provides individualized feedback aimed at improving understanding and expression.

Vocabulary memorization is supported by the *Flashcards page*, which automatically generates review cards from interactions with Harald. Each card contains a Swedish term, its English translation, and metadata identifying its source, enabling targeted review of previously encountered material. Learners can visit this page to practice words Harald has found the user struggling with.

With user data persistently stored in the session state and the modular structure of the system's prompts and conversation history, Harald's interactions gradually evolve to match the learner's progress and preferences,. Over time, this accumulation of contextual information enables precise adaptation of responses and pedagogical focus, reinforcing Harald's role as an attentive and context-aware tutor.

3.1 Previous Exploration

Throughout the project, several language models and large language models (LMs/LLMs) with parameter counts ranging from 3 to 8 billion were evaluated, fine-tuned, and tested. Although the use of LoRA (Low-Rank Adaptation) allows for efficient fine-tuning on domain-specific datasets aligned with the project´s objectives, the overall performance and complexity of Google´s Gemini 2.5 model proved difficult to match. Consequently, Gemini 2.5 Flash was selected as the LLM to build this app on, due to its robustness, conversational quality and adherence to customized system prompts, which ensured adaptability and coherence. Earlier experimental implementations are available in the project´s GitHub repository.

4 Results

4.1 Quantitative analysis

To quantitatively analyze the accuracy and correctness of the responses generated by Harald, a list of 100 questions related to Swedish words and grammar were created. We fed these questions to Harald one at a time to generate a response. The responses were then manually checked for correctness. A four-valued scale was used to label each response:

- **Correct**: the response was correct and complete.
- **Incomplete**: the response was correct, but missed some important details, such as verb gender.
- **Partially incorrect**: the response contained some inaccurate information, such as grammar errors in examples.
- **Incorrect**: the response contained mostly or only incorrect information.

Figure 1 shows the results of the evaluation. As can be seen from the chart, the responses were correct and complete 74% of the time, and partially incorrect 4% of the time. None of the responses were mostly or fully incorrect.

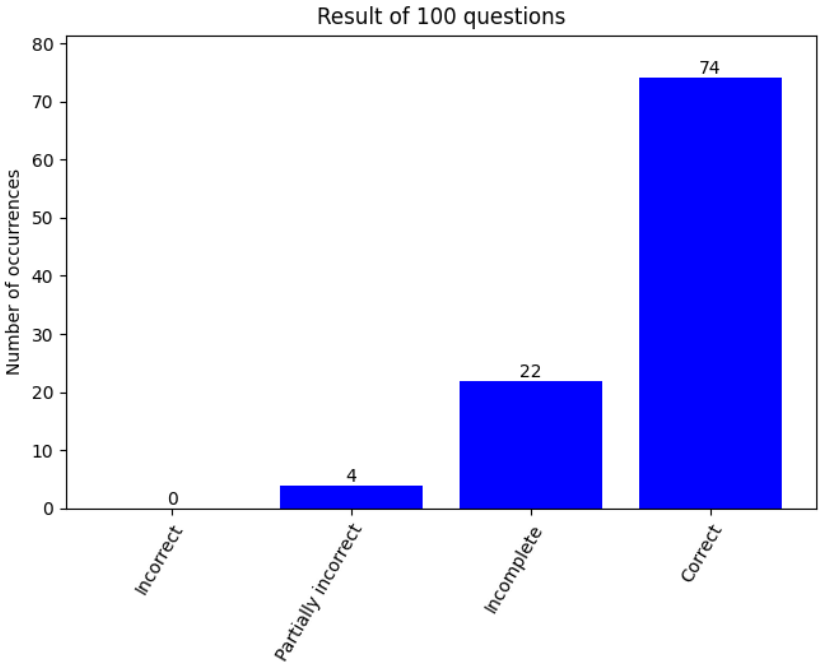

Figure 1: Bar chart showing the quantified accuracy scores.

4.2 Qualitative Analysis

To understand how users experienced the chatbot we shared it with several participants and collected their feedback using a Google Form. The participants included from people who have no proficiency in Swedish to people who have intermediate proficiency. The form included questions about usability, personalization and learning experience. Most of the feedback received was positive showing that users found the chatbot engaging and helpful for practicing Swedish reading and writing. Many participants mentioned that they would use it again or recommend it to others.

When asked *"What kind of improvements would make the chatbot feel more personalized to you?"*, a common suggestion was that the chatbot could ask more questions about the user at the beginning or check their current knowledge level. This would help adapt the responses and examples to each user's background and learning pace.

In response to *"What did you dislike or find confusing?"*, one participant mentioned that the reading texts felt too difficult for their level. This feedback highlights a limitation of the current version that it does not adjust the difficulty of reading materials based on the learner's proficiency.

Overall the qualitative feedback suggests that users appreciated the chatbot's friendly tone and learning support but would like more personalization and difficulty adaptation in future versions. These insights will guide further development to make the chatbot more engaging and effective for learners at different stages in their learning.

5 Societal Impact

The use of large language models (such as Gemini 2.5 Flash) has a strong influence on society. These models can bring many benefits but also create new challenges that must be handled responsibly.

5.1 Positive Implications

The main benefit of Harald is improved accessibility and inclusion. Learners from different backgrounds can practice Swedish anytime without the need for expensive language courses or native-speaking tutors. This can support immigrants, students and professionals who want to integrate into Swedish society more easily. The chatbot also improves educational efficiency offering instant corrections, feedback and explanations in simple English. It encourages personalized learning allowing users to progress at their own pace and supports the democratization of knowledge by making language learning resources available to a wider audience.

5.2 Negative Implications

However, there are several potential risks to be addressed. Since Harald uses the Gemini model, it relies on a pre-trained dataset that we do not have any control over. This means potential biases, inaccuracies or cultural insensitivities from the training data can surface in responses. The only adjustable parameter is the system prompt, which helps guide the model's tone and focus but it might not be effective in all edge cases. There are also parameters which restrict harmful and hateful dialogues but inspite of all these techniques it is difficult to guarantee perfect accuracy or fairness in all interactions. There are also privacy concerns because users might share personal details in their messages. This requires careful data handling and protection. Another risk is that users could become too dependent on the chatbot or overconfident on AI-generated responses and use it instead of real conversations with people. In addition using large models like Gemini has an environmental cost, since it requires high energy use for both training and running the model.

5.3 Ethical Considerations and Mitigation

To manage these issues the system should be transparent about being an AI and not a real teacher. Users should understand that it may not always be correct or unbiased. The system prompts should be regularly updated to make the responses more reliable and human reviews can help check for mistakes or bias. Personal information should be anonymized or removed and data storage should follow privacy regulations. Using smaller or optimized versions of the model can also help lower the environmental impact.

While Harald highlights the educational and social benefits of integrating LLMs into language learning, ongoing monitoring is essential. Balancing accessibility and personalization with fairness, accuracy and ethical responsibility will determine the long-term success and trustworthiness of such systems in society.

Use of generative AI

Generative AI was used for formatting the text for writing this report. It was also used to create example chatbot conversation to use as system prompts for making the chatbot accessible and personlized.

References

[1] Google AI. *Gemini API Text Generation Documentation.* Available at: https://ai.google.dev/gemini-api/docs/text-generation

[2] Google AI. *Function calling with the Gemini API.* Available at: https://ai.google.dev/gemini-api/docs/function-calling