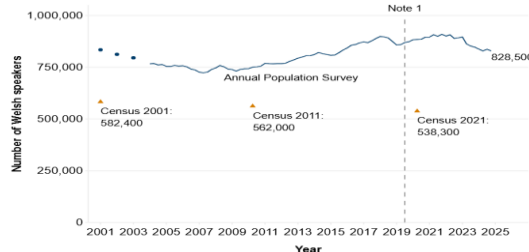


1. Introduction

The Welsh language has always been apart of Welsh tradition and culture, but only approximately 844,300 people speak Welsh today[1], which is about 27.3%.



The Welsh government project “Cymraeg 2050” aims to increase Welsh speakers to 1 million by 2050, with schemes like “Iaith Athrawon Yfory” incentivising students to study teaching through the Welsh medium [2].

With the development of AI increasing and LLMs producing high quality results within larger languages, the project aims to answer whether these same techniques can be applied for lower-resource languages like Welsh.

The development of this application would allow for users to learn and converse in Welsh, allowing for users to practice daily as only 14% of Welsh speakers speak the language daily [1].

2. Aims and Objectives

The aim of this project is to create an application that uses RAG techniques and uses semantic search of Welsh datasets like CorCenCC and Kwici that provides the user with a high-quality answer, reducing hallucinations within the response.

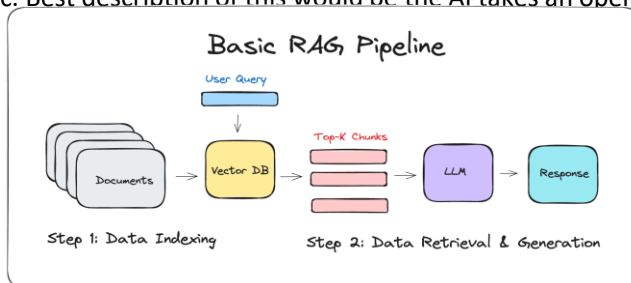
- Analyse and evaluate Welsh literature and datasets to create an intelligent AI tutor that can create sentences in both Welsh and English.
- Develop an application with Dart as it’s front-end UI and a FastAPI in Python that bridges the connection between the UI, the AI and the Welsh datasets using a RAG pipeline.
- Allow the users to determine their proficiency level and the AI provide lessons and responses based on text input by the user.

References

- [1]Welsh Government Language Survey - <https://www.gov.wales/welsh-language-data-annual-population-survey-october-2024-september-2025.html>
- [2] "Iaith Athrawon Yfory Incentive Scheme: guidance for students 2025 to 2026 [HTML] | GOV.WALES." Accessed: May 11 2026. [Online]. Available: <https://www.gov.wales/iaith-athrawon-yfory-incentive-scheme-guidance-students-2025-2026.html>
- [3] I. Belcic. "What is RAG (Retrieval Augmented Generation)? | IBM." Accessed: May 11 2026. [Online]. Available: <https://www.ibm.com/think/topics/retrieval-augmented-generation>

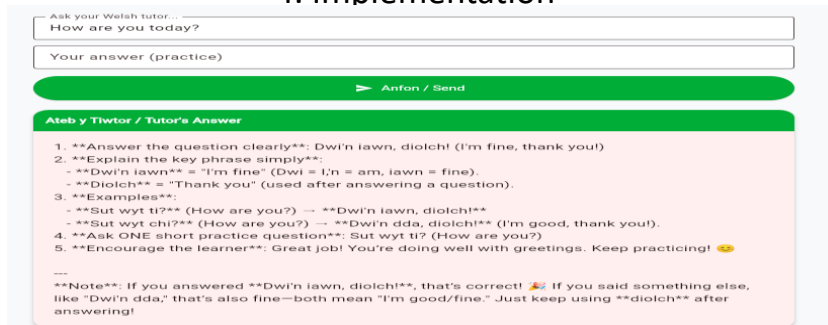
3. What is RAG?

RAG is Retrieval-Augmented Generation [3] is a pipeline that optimises performance of an AI by providing the AI with external databases, which reduces the chance of errors (hallucinations) and removes the need to re-train the AI on a specific topic. Best description of this would be the AI takes an open-book exam



The AI needs to be able to understand the prompt coming in and then compare it to the external dataset. The Python RAG pipeline first chunks the user input (turns the data into vector information) and then compares the chunked input to the external data which has also been chunked. The best response is then chosen and provided to the user

4. Implementation



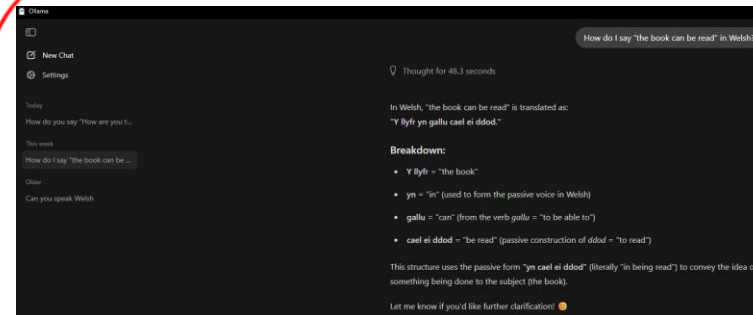
The application was created in Flutter using Dart, which allowed for creation of the UI. This allows user to input-text and the AI will return an answer. The Ollama AI was also provided a prompt which would tell the AI how to respond – like a tutor.

CorCenCC [4] database – a collection of Welsh words and lexicons that were collected from Welsh sources such as journals, web pages and books. The dataset was chunked which allowed for the AI to search semantically to provide the most appropriate response. CorCenCC provides lexicons which reduces errors as the Welsh language has different regional words for the same thing (llefrith and llaeth).

Kwici dataset from Kaggle was also used which provided sentence building, that reduced hallucinations when the AI would build a sentence as English and Welsh differ (Subject-Verb-Object vs Welsh Verb-Subject-Object). Lessons following the CEFR were also created, allowing for more personal tutoring.

- [4] CorCenCc - Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., Muralidaran, V., Tovey-Walsh, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M. and Scannell, K. (2020). CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh. Cardiff University.

5. Conclusion



The image above shows the Ollama LLM being used to prompt the AI to say “the book can be read” in Welsh. It took 48.3 seconds on this specific model.



The Welsh tutor application provides the response much quicker than the original Ollama model has reduced hallucinations as the sentence from Ollama is not correct. Multiple tests conducted show that RAG has been implemented and is a brilliant way to introduce smaller languages to LLMs, provided that there is plenty of online datasets and documents to implement.

6. Future Work and Recommendations

This is a MVP (Minimal-Viable-Product), more work would As this be necessary to create a more retaining application. Gamification would be implemented in the future to improve user retention.

As Welsh is a language, voice input would be researched and implemented, similar to Macsen from Bangor University, which would allow the user to vocally speak Welsh. A conversational part of this application would be implemented in the future, which would allow users to practice speaking Welsh, rather than this application just being a learning application, which is it at the moment. Implementing a login for the application and storing previous conversations would also be implemented, allowing for users to login from a different device and continue the learning/conversation