# Project Proposal: Q&A Chrome Extension

Ye Joon Han, Susie Li, Zhiheng Ye

## Subject/Content Matter

Our project aims to develop a **Q&A Chrome Extension** that can answer users' questions based on the content of the webpage they are viewing. The goal is to reduce the time users spend browsing pages and improve efficiency in finding specific information.

## Data for Fine-Tuning

We plan to fine-tune our model using the **WebGLM-QA dataset**, which is publicly available at Hugging Face. This dataset consists of **question-answer pairs** derived from web content, making it well-suited for our application.

## Data Size

The WebGLM-QA dataset contains **44,979** question-answer pairs extracted from web pages, ensuring a diverse and comprehensive training set, specifically:

- **Train set**: **43,579** rows
- **Validation set**: **1,000** rows
- **Test set**: **400** rows

The total storage size of the dataset is approximately **71MB**, making it lightweight and feasible for fine-tuning while maintaining efficient data processing.

## Potential Biases in the Data

There may be biases in the dataset due to:

- **Source bias**: Web content may not be representative of all perspectives.
- **Answer reliability**: Some question-answer pairs may be generated or extracted with inherent biases.
- **Domain-specific skew**: The dataset may contain more data from certain domains than others, leading to uneven performance across different topics.

Being aware of these biases, we will implement techniques such as dataset balancing and adversarial training to mitigate their impact where possible.

## Pre-trained Model

We will use **DeepSeek-R1-Distill-Llama-8B**, a distilled version of **Llama-8B**, as our base model for fine-tuning. This model is well-suited for **natural language understanding and generation**, making it a strong candidate for our Q&A application.

## Outcome/Deliverable

The final deliverable will be a **Chrome Extension** that:

- Extracts relevant content from the webpage.
- Accepts user queries and provides accurate answers based on the webpage context.
- Integrates seamlessly into the browser, offering an intuitive and user-friendly experience.