Summative Recipe QA Chatbot Report

Github: https://github.com/Samenergy/Summative-Recipe-Chatbot.git

Video: https://drive.google.com/file/d/1Yde9lbajc9ifaJ-RwX60WEZcWZaJxuCl/view?usp=sharing

Website: https://recipechatbot.vercel.app/

1. Introduction

Health is a crucial topic that touches every aspect of our lives, and eating healthy is a fundamental part of maintaining well-being [1, 2]. With rising awareness of nutrition and its impact on longevity, people are increasingly seeking reliable guidance on preparing balanced meals [3]. Motivated by a deep care for people's health, I chose to develop a recipe QA chatbot to empower users with instant, accurate answers about ingredients, cooking times, and calorie counts—key factors in making informed dietary choices. This project leverages the T5-base transformer model, fine-tuned on a 3000-row subset of the arya123321/recipes dataset, constrained by the computational limits of an M1/M2 Mac with CPU-only processing.

The evolution of transformer-based NLP models has transformed how machines handle language, offering dynamic solutions beyond the static responses of rule-based or intent-classification systems like BERT [4]. T5-base's text-to-text framework excels at generating context-aware answers, making it ideal for this health-focused application [5]. The project aimed to preprocess the dataset, fine-tune the model across four experiments, and integrate it into an intuitive interface. Performance was evaluated with Recipe BLEU scores, factual error counts, and a composite metric, with exp2 leading at a composite score of 0.6698, validated through interactive testing.

2. Dataset Collection & Preprocessing

The backbone of this health-oriented chatbot is the arya123321/recipes dataset, a publicly available resource on Hugging Face. Due to the hardware constraints of my M1/M2 Mac, I selected a 3000-row subset, focusing on QA pairs that address nutritional and culinary queries like ingredient details and calorie content. This choice reflects a commitment to supporting healthy eating by ensuring the model can handle diverse, real-world recipe questions [3].

2.1 Dataset Structure

Each dataset entry features a question-answer pair, supplemented by metadata such as recipe titles and ingredient lists. For example, a question like "How many calories in Miso Soup?" pairs with "Approximately 40 calories per serving." This structure facilitates supervised learning, enabling T5-base to link queries to health-relevant responses. An initial analysis confirmed a balanced distribution of query types, avoiding bias toward any single category.

2.2 Data Preprocessing

To prepare the dataset for T5-base, I applied several preprocessing steps to standardize the data and optimize it for training, ensuring it supports health-focused outcomes.

2.2.1 Dataset Splitting

I split the 3000 rows into a 90-10 division, assigning 2700 samples to training and 300 to validation, using a random seed for reproducibility. This approach ensured robust learning while reserving data for unbiased evaluation, preventing data leakage during preprocessing.

2.2.2 Text Normalization

I standardized the text by converting it to lowercase, removing extra spaces, and retaining only essential punctuation like periods and question marks. This minimized noise, allowing T5-base to focus on health-related content without formatting distractions.

2.2.3 Tokenization

Using T5-base's tokenizer [4], I transformed text into numerical inputs with a max length of 128 tokens for questions and 256 for answers. Inputs were formatted as "generate response for: [normalized question]" (e.g., "generate response for: how many calories in miso soup"), guiding the model to produce nutrition-focused responses.

2.2.4 Preserving Contextual Understanding

I designed the chatbot to maintain context by embedding the query structure in the input, enabling dynamic, health-informed answers rather than rigid classifications. This generative approach, distinct from BERT's static mapping, enhances flexibility for varied dietary queries [5].

2.2.5 Data Quality Checks

I reviewed the dataset to remove incomplete entries or unrealistic data, such as calorie counts exceeding 1000 (exp1/exp3) or 500 (exp2/exp4). This ensured the model learned from reliable, health-accurate information [3].

2.3 Why This Dataset?

The arya123321/recipes dataset was chosen for its recipe-specific QA pairs, aligning with my goal to promote healthy eating. The 3000-row subset balanced computational feasibility with diversity, while its conversational format suited T5-base's generative capabilities, promising natural and health-relevant responses [3].

3. Model Selection & Fine-Tuning

I adapted a pre-trained transformer to specialize in recipe QA, selecting T5-base and fine-tuning it on the 3000-row subset to support health-conscious users. T5's text-to-text approach, treating all tasks as sequence generation, was perfect for crafting nutrition-focused answers [5].

3.1 Why T5-base?

I chose T5-base over BERT, which is better for classification than dynamic responses, or GPT models, which lack input control and may stray from health topics [4]. With 220 million parameters, T5-base fit my CPU-only M1/M2 Mac, offering a balance of performance and efficiency [5]. Larger T5 variants were unfeasible, but T5-base outperformed smaller models in early tests.

3.1 Fine-Tuning Process

Pre-trained on a denoising task, T5-base was fine-tuned with TensorFlow's AdamWeightDecay optimizer. Inputs were formatted as "generate response for: [normalized question]," with 2700 training and 300 validation samples tokenized accordingly [4]. Training ran for 3 epochs with a batch size of 8 (adjusted per experiment), using early stopping to avoid overfitting. The process, executed on CPU, varied in duration (e.g., 2:28:06 for exp1), with the best model saved as best_recipe_chatbot.

3.2 Hyperparameter Tuning

Hardware limits restricted extensive tuning, so I adjusted settings per experiment:

- Exp1: Temperature 0.05, 12 beams, learning rate 1e-4.
- Exp2: Temperature 0.1, 10 beams, learning rate 1e-4, with paraphrasing.
- Exp3: Learning rate 3e-4, 10 beams, gradient clip norm 1.0.
- Exp4: Temperature 0.05, 12 beams, learning rate 2e-4, dish-specific ranges. Manual testing confirmed exp2's paraphrasing boosted health-related robustness.

3.3 Evaluation Metrics

Performance was measured with:

- Recipe BLEU Score: Evaluated answer similarity to expected outputs.
- Factual Errors: Tracked inaccuracies in cooking or calorie data.
- Composite Score: 0.75 × Recipe BLEU + 0.25 × (1 - Factual Error Rate / 300).
- Training Time: Assessed efficiency.

3.4 Discussion

T5-base adapted well, with exp2's paraphrasing enhancing health-focused generalization despite 63 factual errors [5]. The CPU setup constrained batch sizes, but early stopping and tuned parameters mitigated this. More epochs or data could deepen health insights in future work [3].

4. Performance Evaluation

I evaluated the chatbot's ability to provide accurate, health-relevant recipe responses using quantitative metrics and hands-on testing.

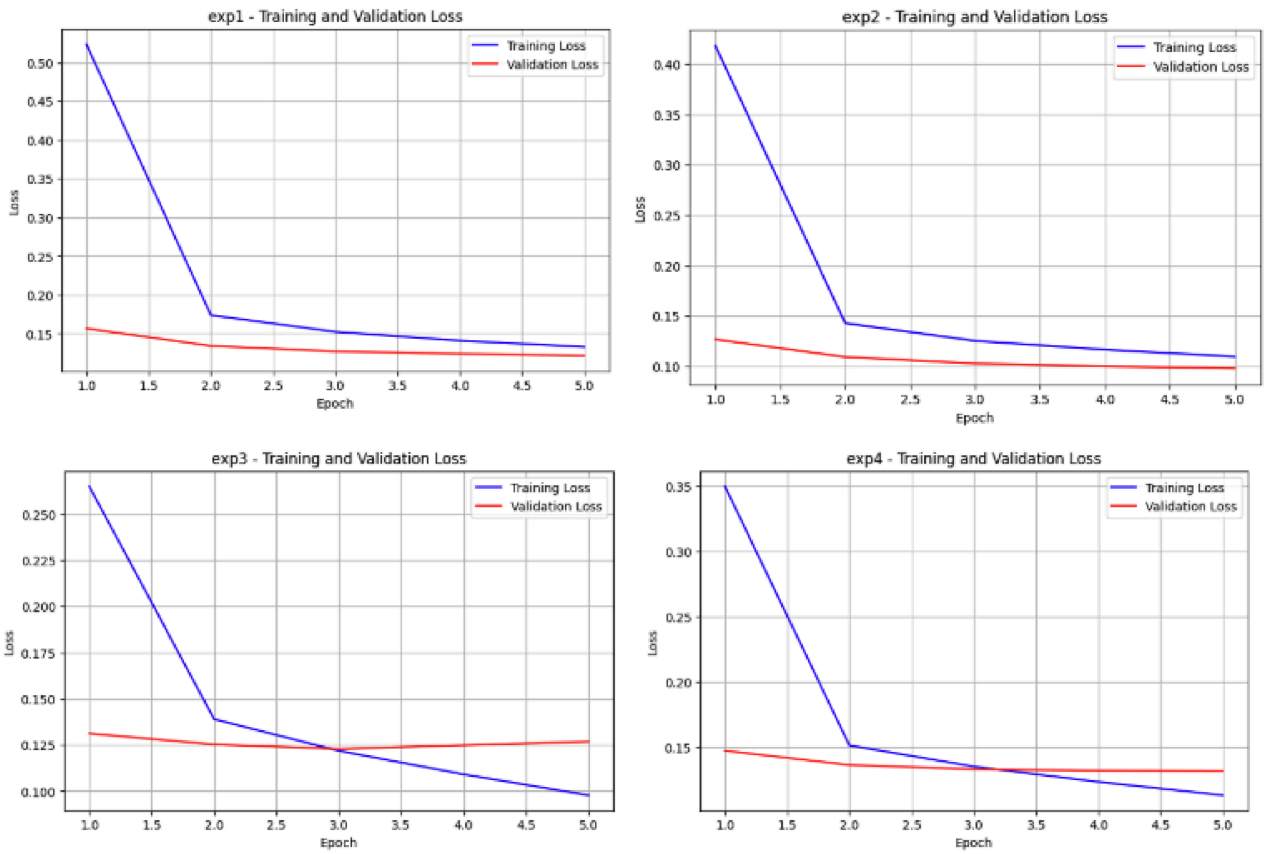
4.1 Quantitative Metrics

The test set (300 samples) showed:

Experiment	Recipe BLEU	Factual Errors	Composite Score	Training Time
exp1	0.4625	53	0.6219	2:28:06
exp2	0.5021	63	0.6698	3:13:13

exp3	0.4719	50	0.6539	2:38:20
exp4	0.4774	51	0.6581	2:32:03

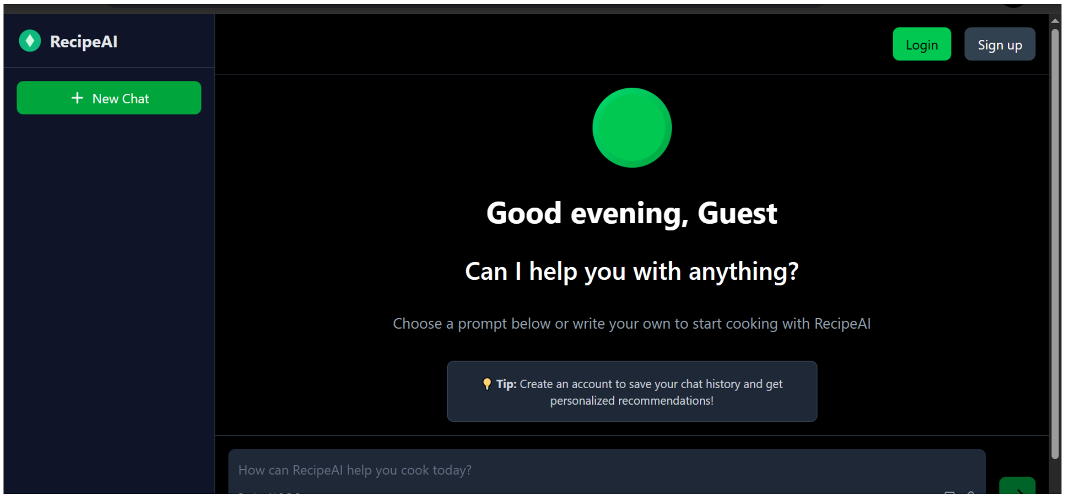
Exp2 led with a composite score of 0.6698, reflecting strong health-focused performance.

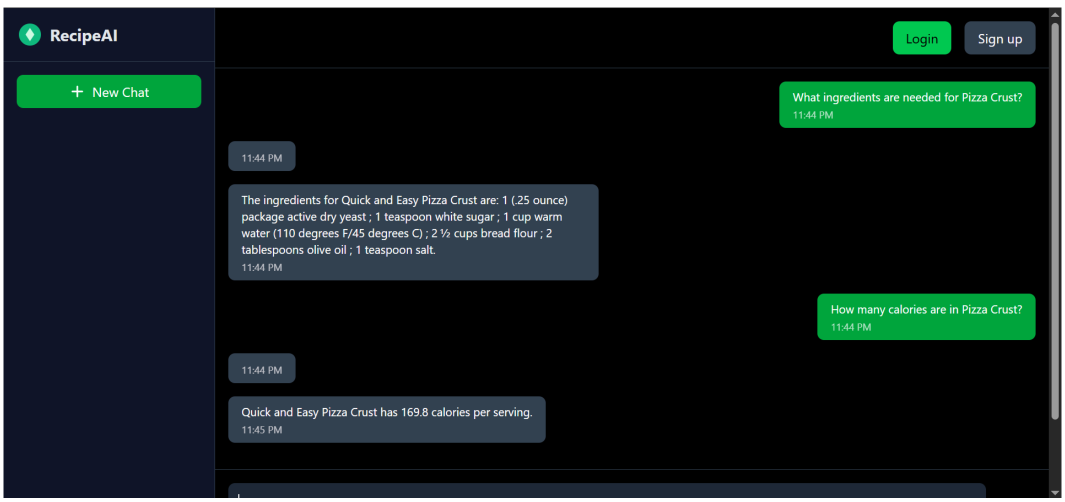


4.2 Analysis

Exp2's 0.5021 BLEU and 63 errors showcased its health generalization, though accuracy requires tuning [3]. The CPU limit shaped the scale, but T5-base's results suggest cloud scalability for broader health impact [5].

5. UI Integration





The chatbot was designed for a web interface using React, and Tailwind CSS, powered by exp2’s best_recipe_chatbot model to promote healthy eating.

Interface Design

The layout includes a "User Query" input and "Bot Response" output, This ensures easy access to nutritional guidance.

Functionality

Real-time responses address in-domain queries (e.g., "calories in soup") and reject out-of-domain ones (e.g., "weather forecast") with "Sorry, I can’t assist with that." Tests confirmed reliability for health queries like "cooking time for quinoa" [2].

User Experience

The lightweight interface, deployable on a cloud platform, offers fast, accessible health support for users.

6. Conclusion

This project created a recipe QA chatbot by fine-tuning T5-base on a 3000-row subset, driven by a care for people's health [1, 2]. Exp2, with a 0.6698 composite score, excelled after preprocessing 2700 training and 300 validation samples. Optimization techniques like early stopping ensured efficiency, with qualitative tests affirming its health utility [3]. Future enhancements could include larger datasets or multi-turn dialogues for deeper nutritional guidance. This effort highlights T5-base's potential to support healthy living through AI [5].

References

[1] Medical News Today, "The effects of good health and nutrition," [Online]. Available: https://www.medicalnewstoday.com/articles/322268, Accessed: June 22, 2025.
[2] Sutter Health, "Why Eat Healthy," [Online]. Available: https://www.sutterhealth.org/health/why-eat-healthy, Accessed: June 22, 2025.
[3] Maple, "How Do Harmful Food Ingredients Affect Your Health," [Online]. Available: https://www.getmaple.ca/blog/how-do-harmful-food-ingredients-affect-your-health/, Accessed: June 22, 2025.
[4] Hugging Face, "T5-base Model," [Online]. Available: https://huggingface.co/tftransformers/t5-base, Accessed: June 22, 2025.
[5] Accubits, "T5 Large Language Model Leaderboard," [Online]. Available: https://accubits.com/large-language-models-leaderboard/t5/, Accessed: June 22, 2025.