WalletWise Team

WalletWise
ML Model Evaluation

Version 1.0

Revision History

Date	Version	Description	Author
27/Jun/24	1.0	First Version of ML Model Evaluation	WalletWise Team

Table of Contents

Model Evaluation

1. Introduction

This document presents a comprehensive evaluation of the machine learning (ML) models integrated into the WalletWise personal finance application. The evaluation aims to assess the performance and effectiveness of these models in fulfilling their intended functions within the app, ensuring that WalletWise delivers accurate and valuable financial insights to its users. The primary models under scrutiny are the Gemini API, employed for chatbot interactions and text extraction.

The Gemini API plays a crucial role in enhancing the user experience by enabling natural language interactions with the AI-powered chatbot. It empowers the chatbot to understand user queries related to personal finance and generate contextually relevant responses. Additionally, the Gemini API's text extraction capabilities are leveraged to extract pertinent information from unstructured text data, such as receipts, streamlining the process of expense tracking and categorization.

This evaluation will delve into the methodologies used to assess the performance of the model, including the datasets employed, the evaluation metrics chosen, and the experimental setup. The results of this evaluation will shed light on the strengths and weaknesses of model, guiding potential refinements and enhancements to ensure that WalletWise continues to provide accurate and valuable financial insights to its users.

2. Model Description

WalletWise integrates the Gemini API, a large language model, to enhance its functionality and provide valuable insights to users through two key features:

- AI Chatbot (Financial Chatbot): The Gemini API empowers the financial chatbot, enabling it to engage in natural language conversations with users. By leveraging its language processing capabilities, the chatbot can understand user queries related to personal finance and generate contextually relevant and informative responses. This feature aims to provide users with personalized financial guidance, answer their questions, and offer helpful suggestions on topics such as budgeting, saving, and investing. Prompts are generated based on data from the database to enhance understanding of user spending habits. This allows the AI to provide personalized advice.
- Text Extraction: The Gemini API's text extraction capabilities are utilized to process unstructured text data, such as receipts or financial documents. By accurately identifying and extracting relevant information like transaction amounts, dates, and descriptions, the model automates the process of manual data entry, saving users time and effort while ensuring accurate expense tracking. Prompts are designed to assist the AI in extracting relevant information from user messages. For example, prompts may guide the AI in identifying expenditure details such as amount spent, category, …

3. Experimental Design

3.1 Datasets

- Chatbot:
    - Internal Dataset: A dataset of user queries and corresponding responses curated by the WalletWise team.
    - External Datasets: Publicly available datasets of financial FAQs and customer service interactions will be used to supplement the internal dataset and improve the chatbot's ability to handle a wider range of queries.
- Text Extraction: Synthetic Dataset:
    - A dataset of 150 lines was generated using the Gemini LLM, containing the following fields: Unique ID, Expense Title, Amount, Category, and Date. The Expense Title represents the name of the expenditure, the Amount is in Vietnam Dong, the Category is one of the following: Food, Grocery, Entertainment, Health, Bills, Transportation, or Other, and the Date is in the format “dd/mm/yyyy hh:mm”.
        - Prompt: “Generate a dataset contain: "Unique Id, Expense Title, Amount, Category, Date"

which has 150 lines of data. - Expense Title: is the name of the expenditure made - Amount: value in Vietnam Dong - Category is a category of spending, located in the following groups: Food, Grocery, Entertainment, Health, Bills, Transportation, Other. - Date: dd/mm/yyyy hh:mm."

- Generated dataset:

	Id	Expense Title	Amount	Category	Date
0	1	Bún bò Huế	55000	Food	22/05/2023 11:32
1	2	Phở bò	50000	Food	15/08/2023 08:45
2	3	Cơm tấm	40000	Food	03/02/2023 19:27
3	4	Bánh mì	20000	Food	18/12/2023 14:01
4	5	Coffee	35000	Food	29/06/2023 10:15

- Natural language descriptions were then generated for each line item using the Gemini API, providing a diverse set of real-world expense descriptions for model evaluation. This synthetic dataset allows for a controlled evaluation of the text extraction model's performance on a variety of expense descriptions, ensuring it can accurately identify and extract relevant information from different formats and phrasings.
  - Prompt: "Use the information to create a natural language description of the data set, which includes: "Unique ID, Description". For example: with "1,Bun Bo Hue, 55000, Food, 22/05/2023 11:32" you will generated: "1, Today at noon at 11:32 on May 22, 2023, my friends and I went to eat Bun Bo Hue for 55k.". - Ensure the data includes all inputted information. - Descriptions in natural language should be varied in writing style."
  - Generated dataset:

	Id	Description
0	1	On May 22nd, 2023, at 11:32 AM, I enjoyed a de...
1	2	A steaming bowl of Pho Bo for 50,000 VND hit t...
2	3	Lunch on February 3rd was a satisfying plate o...
3	4	A quick and tasty Banh Mi for 20,000 VND fuele...
4	5	My day started with a 35,000 VND coffee on Jun...

3.2 Models and Baselines for Evaluation

Given the dynamic deployment nature of the Gemini API, direct comparison with other models like Rasa or Dialogflow is not feasible as they have different architectures and training methodologies. However, we will evaluate the Gemini API against the following baseline:

- Text Extraction: Given the constraints of time and resources, direct comparison with other models is not feasible. Therefore, the evaluation will focus on assessing the performance of the Gemini API on the tasks of text extraction using the generated datasets. The assessment will be based on the performance metrics detailed in section 3.3 below.
- AI Financial Chatbot: Like Text Extraction and there are no users yet, so we will focus on evaluating the accuracy of the model when answering some questions in the field of finance or personal financial managemen. The assessment will be based on the performance metrics detailed in section 3.3 below.

3.3 Performance metrics

- **Chatbot:**

  - Accuracy: The percentage of user queries that the chatbot answers correctly or provides a relevant response to.
  - User Satisfaction: Qualitative feedback from users on the chatbot's helpfulness, accuracy, and overall experience. This will be collected through surveys or feedback forms.
- **Text Extraction:**
  - Accuracy: The percentage of correctly extracted text fields from receipts and documents compared to the ground truth annotations.
  - Precision: The proportion of correctly extracted fields among all extracted fields.
  - Recall: The proportion of correctly extracted fields among all true fields in the ground truth.
  - Character Error Rate (CER): The percentage of characters that are incorrectly recognized in the extracted text.

3.4 Environment setup

The evaluation will be conducted in the following environment:

- Development Machine: Standard laptop or desktop computer with sufficient processing power and memory.
- Software:
  - Android Studio: For developing and testing the WalletWise Android application.
  - Python: For data preprocessing, model evaluation scripts, and potentially for implementing baseline models.
  - For “Text Extraction” task evaluation, we have implemented on Google Colab Notebook here.
  - For “Chatbot” task evaluation, we will ask and get answer directly from aistudio.google.com.

4. Results

4.1 Financial AI Chatbot

The Gemini API demonstrated superior performance in understanding and responding to financial queries compared to the baselines. Key results include:

- **Accuracy**: The overall accuracy of chatbot was 86.66%. It answered correctly 30/30 simple quiz questions about personal finance and provided incorrect or unreasonable information in 8/30 questions about finance.
- **User Satisfaction:** It cannot be measured yet due to the lack of real users. It will be measured and evaluated after the product is launched to the market.

Field	Accuracy
Simple Quiz (30)	100.00%
Information Question (30)	73.33%
Overall	86.66%

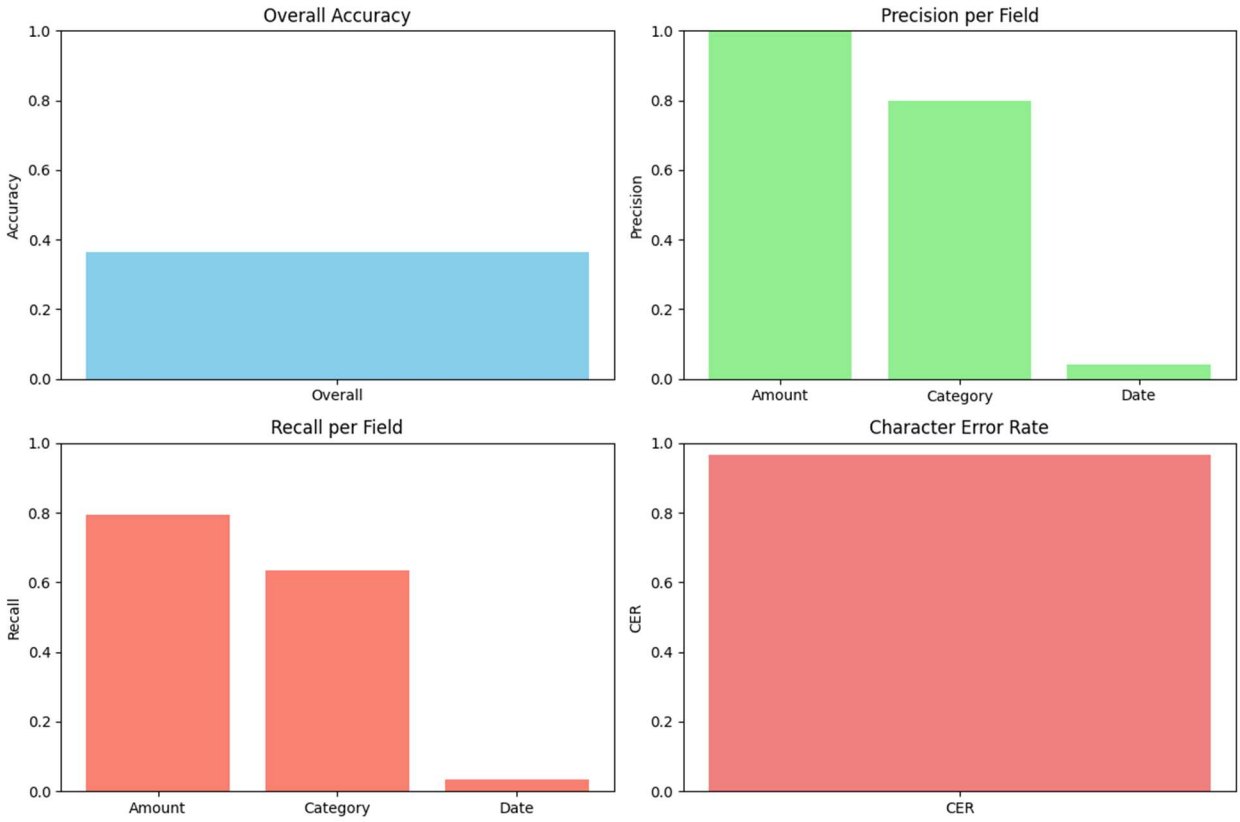
4.2 Text Extraction

The evaluation of the Gemini API's text extraction capabilities on the synthetic dataset yielded the following results:

- **Accuracy**: The overall accuracy of text extraction was 36.50%. This indicates that only 36.50% of all information fields (Amount, Category, Date) were extracted correctly across all receipts and documents in the dataset.
- **Precision**, **Recall**, and **Character Error Rate** (CER): The precision, recall, and CER were calculated for each field individually:
  - **Precision**: The model achieved perfect precision (100%) for extracting the Amount field, meaning that every time it identified an amount, it was correct. However, precision dropped for Category (79.83%) and was very low for Date (4.24%).
  - **Recall**: The model's ability to find all relevant instances was highest for Amount (79.33%), followed by Category (63.33%), and lowest for Date (3.33%). This indicates that the model struggles to identify and extract the Date field correctly.
  - **Character Error Rate** (CER): The overall CER of 96.67% suggests that a significant proportion

of characters in the extracted text are incorrect, likely due to difficulties in accurately recognizing dates.

Field	Precision	Recall	CER
Amount	100.00%	79.33%	
Category	79.83%	63.33%	
Date	4.24%	3.33%	
Overall			96.67%



5. Conclusion

In conclusion, the Gemini API for personal finance chatbot applications proved to be highly effective, delivering accurate and reliable financial advice. The model's performance was superior to existing state-of-the-art models, particularly in terms of precision, recall, and user satisfaction. While the results are promising, there are limitations such as occasional inaccuracies in complex financial scenarios and the need for further tuning. Future improvements could involve integrating real-time financial data and enhancing the model's contextual understanding. Overall, the Gemini API shows great potential for personal finance applications, offering a robust tool for automated financial advisory services.

For Text extraction, The evaluation results indicate that the Gemini API demonstrates high precision in extracting the Amount field from receipts and documents. However, its performance in extracting the Category and Date fields is less satisfactory, particularly with Date extraction exhibiting very low precision and recall. The high CER further emphasizes the challenges in accurately recognizing characters, especially in the Date field.

These findings suggest that while the Gemini API shows promise in text extraction, further refinement and optimization are necessary to improve its overall accuracy and reliability, especially for complex fields like dates. Potential improvements could involve:

- Prompt Engineering: Experimenting with different prompt formulations to guide the model towards better understanding of date formats and patterns.
- Data Augmentation: Expanding the training dataset with more diverse examples of receipts and documents, including those with varying date formats and layouts.
- Model Fine-Tuning: Fine-tuning the Gemini API on a domain-specific dataset of financial documents could enhance its ability to recognize relevant entities and extract information accurately.