Envitas Sprint Review #3

Jeffrey, Mike, Gabriel, Cho, Topher

Overall Project Update

- Progress is moving as anticipated
    - Currently we have three main divisions of work:
        - Frontend/UI
        - Llamasearch
        - Backend/Data
    - In the coming weeks we will begin integrating this split

Response to Feedback

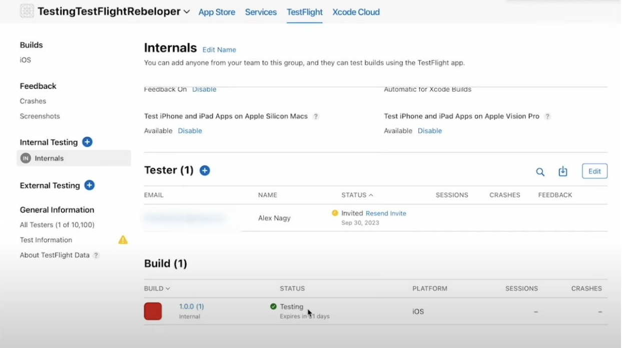
- Security concerns
  - The LLM will have read only access to the public restaurant database
    - This means only select statements can be run
    - Sensitive user information will not be available
  - We can add another layer of security against prompt injection with LLM Guard
- LLM Accuracy/Efficiency
- Novelty of app when compared to current GPT/Gemini

Response to Feedback

- Additionally, we can use AWS add-ons for security on the data side. (Aiven)
    - Point in time recovery of dataset in case of unforeseen breaches
    - Read-replicas for disasters
    - Running maintenance checks, and installing security patches
    - Access logs

Mike Sprint Update

- Created dummy compliance report to upload project to Testflight for demo testing
- added dummy project to test on personal devices



Internals

You can add anyone from your team to this group, and they can test builds using the TestFlight app.

Feedback On Disable | Automatic for Xcode Builds

Test iPhone and iPad Apps on Apple Silicon Macs ? Available Disable

Test iPhone and iPad Apps on Apple Vision Pro ? Available Disable

Tester (1)

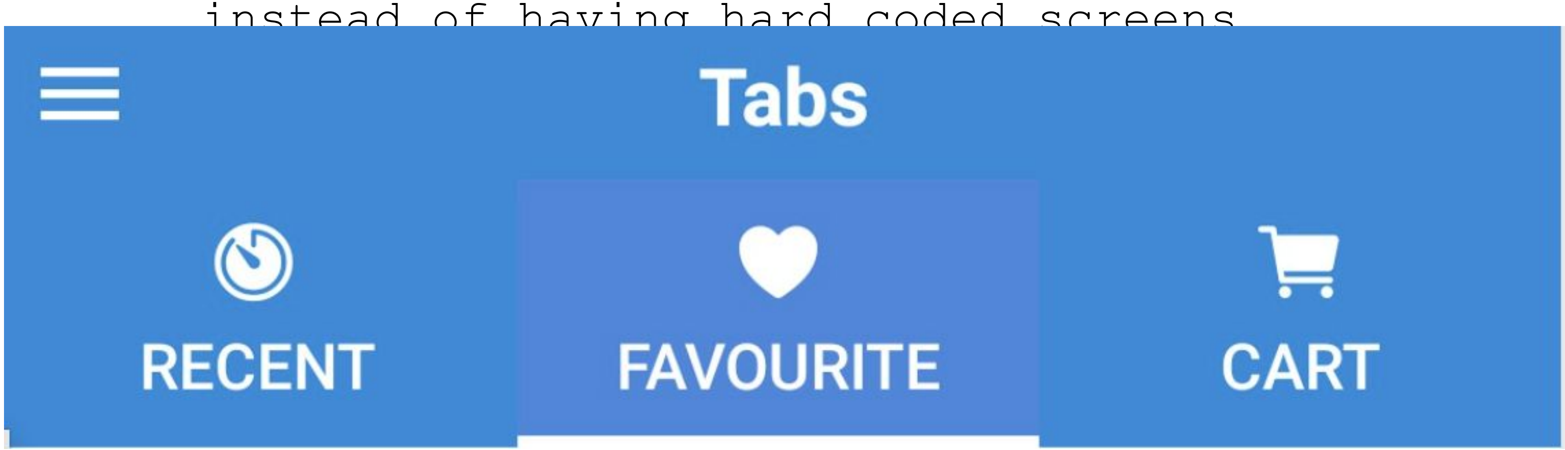
EMAIL	NAME	STATUS ^	SESSIONS	CRASHES	FEEDBACK
[redacted]	Alex Nagy	Invited Resend Invite Sep 30, 2023			

Build (1)

BUILD	STATUS	PLATFORM	SESSIONS	CRASHES
1.0.0 (1) Internal	✓ Testing Expires in 81 days	iOS	–	–

Mike Sprint Update

- added switch between tabs
- removed redundant headers which switching between tabs
-tested different tab layouts for design
-refactored code to have clear routes instead of having hard coded screens



Gabriel Sprint Update:

Previous:

- Basic Dataset Filtering using LlamaIndex - Done
- Use RAG to rerank remaining restaurants - Trouble

Fix/To-DO:

- Do a single pass of our review dataset on an LLM for it to score different aspects of each review (review gives an 8/10 score for food, 5/10 for drinks, and 10/10 for ambiance, for ex.)
- Now our review dataset is structured and socored - Use those scores to rerank the restaurants.

Gabriel Sprint Update:

Concerns:

- Might be a high cost to do this singular pass on our review set, as it is extensive.
- Verify if using RAG to get the most relevant review to the search query is still viable.
-

Search Costs/Time

API COSTS:

GPT 4-Turbo ~ $0.01 per search
GPT 3.5-Turbo ~ $0.01 per 25 searches

GPT 4-Turbo ~ $1 per 100 searches
GPT 3.5-Turbo ~ $1 per 2500 searches

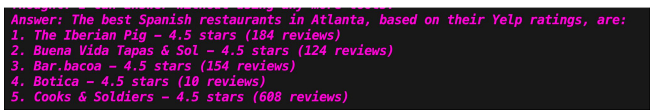
Open Source Model: Possibly cheaper (Pay per GPU usage time)

Runtime:

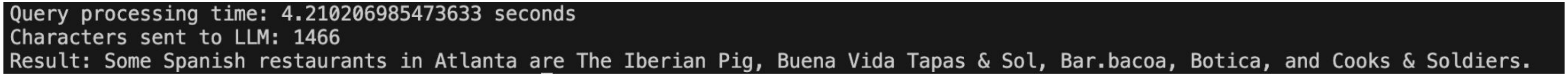
GPT 4-Turbo ~ 18s

GPT 3.5-Turbo ~ 4s

OBS: Only problem with GPT 3.5 is suboptimal structure of final response



```
Thought: I can answer without using any more tools.
Answer: The best Spanish restaurants in Atlanta, based on their Yelp ratings, are:
1. The Iberian Pig - 4.5 stars (184 reviews)
2. Buena Vida Tapas & Sol - 4.5 stars (124 reviews)
3. Bar.bacoa - 4.5 stars (154 reviews)
4. Botica - 4.5 stars (10 reviews)
5. Cooks & Soldiers - 4.5 stars (608 reviews)
```



```
Query processing time: 4.210206985473633 seconds
Characters sent to LLM: 1466
Result: Some Spanish restaurants in Atlanta are The Iberian Pig, Buena Vida Tapas & Sol, Bar.bacoa, Botica, and Cooks & Soldiers.
```

Jeffrey Sprint Update

- AWS RDS
- AWS IAM
- Schema implemented and data pushed to RDS
    - Data converted from JSON → CSV → SQL
    - Copy data to postgreSQL server
- Feedback – more data info

Next Steps and further considerations

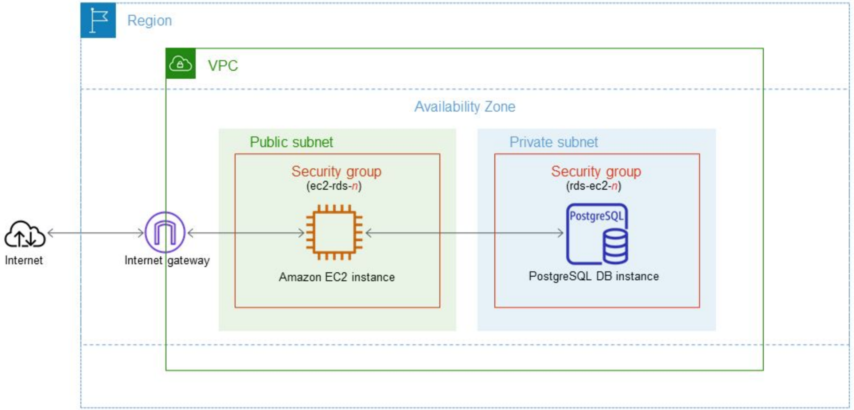
- AWS Bill Management

Topher Sprint Update

- Security concerns and addressing unrelated prompts
- AWS data storage updates

Next Steps and further considerations

- Better AWS architecture
- Connect to frontend



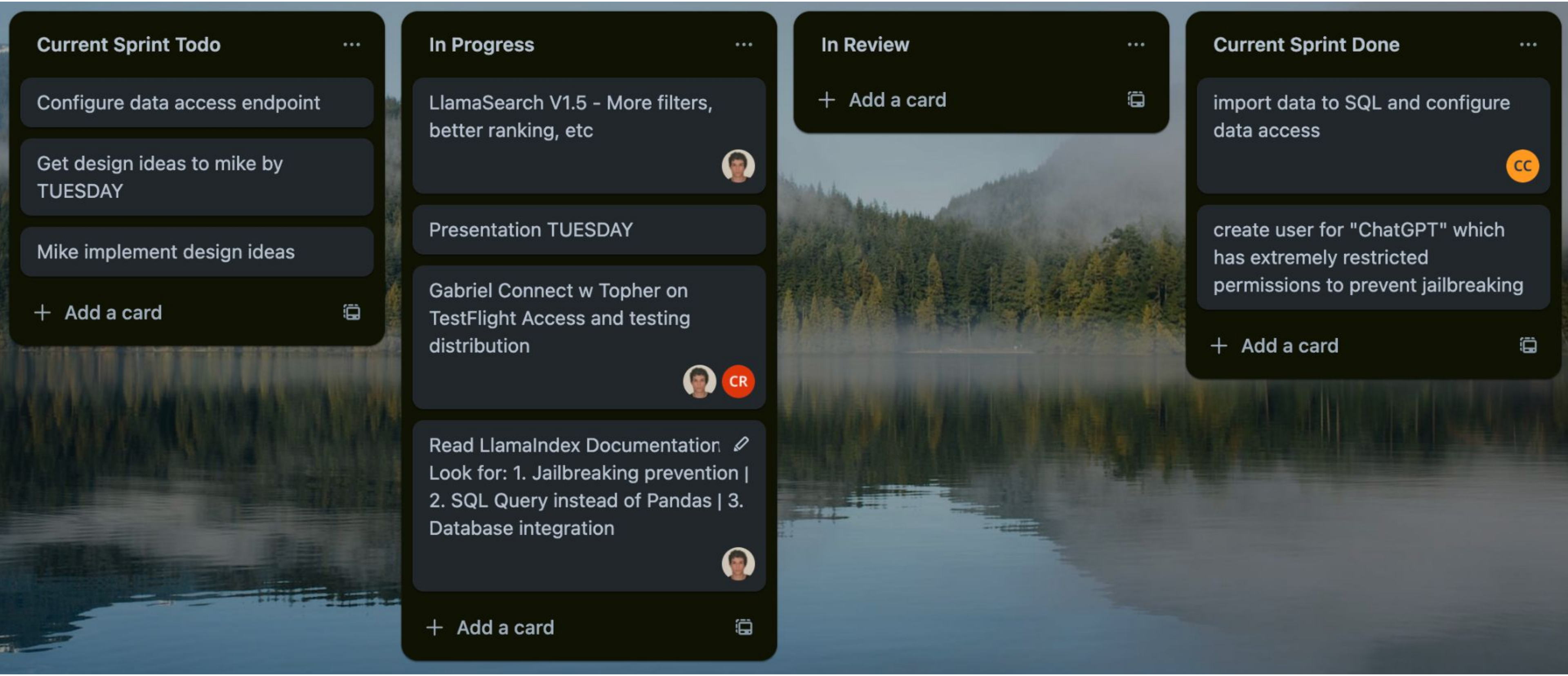
Cho Sprint Update

- Looked into SQL AI Generators
- Looked into security measures specific to dataset
- Running python code on EC2 instances

Next steps

- Running our python code on EC2 instance

Sprint #4 and Deadlines Backlog



Current Sprint Todo

- Configure data access endpoint
- Get design ideas to mike by TUESDAY
- Mike implement design ideas

In Progress

- LlamaSearch V1.5 - More filters, better ranking, etc
- Presentation TUESDAY
- Gabriel Connect w Topher on TestFlight Access and testing distribution
- Read LlamaIndex Documentation Look for: 1. Jailbreaking prevention | 2. SQL Query instead of Pandas | 3. Database integration

In Review

Current Sprint Done

- import data to SQL and configure data access
- create user for "ChatGPT" which has extremely restricted permissions to prevent jailbreaking

Questions?