

# Milestone 3: MinuteMate

Aboli Kasar - Yash Pradhan - Riley LePrell - Neal Logan

## Milestone 2 Completion Status

### 1. PyTest Integration

- Added at Milestone 3
- Integrated unit tests for critical project modules using PyTest, specifically:
  - The Preprocessing Pipeline:
    - Dependencies Check
      - Ensuring all required dependencies are installed.
    - Environmental Variables Check
      - Verifying all API keys/Connection Strings are set up.
    - Azure Upload and Download Verification
    - PDF Conversion Tests
  - The FastAPI backend of the chat application.

### 2. GitHub Workflow Enhancement

- Completed at Milestone 2
- Updated GitHub Actions workflows to include automated testing and code linting on every pull request.
- Added a badge to the repository's README file indicating the build status.

### 3. RAG Implementation

- Completed at Milestone 2
- Implemented a RAG pipeline using a vector database for efficient retrieval.
- Successfully integrated the retrieval module into the LLM interaction flow to improve contextual accuracy.

### 4. Repository Organization

- Completed at Milestone 2, refined at Milestone 3
- Reorganized the repository for clarity and usability:
  - Created separate directories for `src`, `tests`, `docs`, and `data`.
  - Enhanced the README file with detailed setup instructions and examples.

### 5. Project Use Case Definition

- Completed at Milestone 2, refined at Milestone 3
- Refined the use case for MinuteMate as an AI-driven tool for facilitating access to the contents of public meetings.
- Defined primary user personas: researchers, journalists, local government staff, and other interested citizens who want to efficiently access the contents of public meetings.

## Testing Results and Analysis

**Selected Methodology:** Systematic Testing

**Testing Suite:** Implemented with [DeepEval](#) for performance assessment.

## Performance Metrics & Justification

- **Speed Metrics** - Low priority, except that it absolutely can't be extremely slow
  - **Time to First Token** - needs to start promptly, within a second or so, for all responses.
  - **Tokens Per Second** - needs to proceed fast enough to avoid losing engagement for all responses.
  - **Time to Last Token** - this applies differently depending on the type of response. A long time to the last token can mean an excessively verbose response and may interfere with the user's ability to direct the flow of the conversation to keep it relevant.
- **Contextual (Information Retrieval) Accuracy** - These measures are important diagnostics to determine why response accuracy might be suffering.
  - **Precision**
  - **Recall**
  - **Relevance**
- **Response (Contextual Generation) Accuracy**
  - **Relevancy** - Third priority. Irrelevant information wastes people's time and does not provide value.
  - **Faithfulness** - Second priority. The point of the app is to provide information based on the actual contents of public meetings.
  - **Hallucination** - Top priority. Even a fairly low rate of fabricated information could mean that this app provides net harm rather than net benefit.
- **Trust and Safety**
  - **Toxicity - somewhat important** because this is a public website, but even top language models are seeing some egregious failures so the risk

of embarrassment or additional harm is not high. We'll want some checks for toxicity to ensure the app's responses are not egregious.

- **Privacy** - these are public meetings, so in general, privacy concerns are minimal. However, the app should generally refuse to engage in clear privacy violations, such as queries seeking sensitive personal information that might have been introduced into meetings.
- **Security** - We used a security-by-design approach, presenting no interface for the general public to introduce new content into the database. The only way they can interact with the system is through the public Streamlit chat app, or if they find the endpoint, the chat application API. We also use a security linter and will update dependencies to avoid vulnerabilities.

### Impact on Development:

- Early testing results are bringing to light critical problems with core functionality, helping guide further development efforts. In particular, the absence of critical metadata associated with our chunks made information retrieval targeted at specific timeframes or meeting dates mostly impossible. Between the difficulty of retrieval and the lack of date labels on the chunks, the generative model could not properly contextualize the retrieved information, leading it to decline to use even highly relevant retrieved information because it could not be confident in its relevance.

## Final Project Planning

### Timeline for Remaining Development

- **Nov 25th - Dec 5th: Critical Development**
  - Implement critical feature - date metadata for chunks and the ability to filter queries by date. The app is mostly nonfunctional without this.
  - Add effective logging with export capability in useful/analyzable format from staff app in order to implement the virtuous cycle of a data flywheel.
- **Nov 25th - Dec 10th: Targets for Development, Cleanup, and Testing**
  - Enhancements:
    - Add safety/toxicity checks to chat app
    - Add additional metadata
    - Add Llama3.2 11B hosted on Modal
  - Cleanup:

- Remove references to endpoints and keys from public Streamlit app
- Remove excess files, imports, etc. to ensure all linters pass
- Clean up documentation, requirements, dockerfiles, etc.
- Testing:
  - Expand and refine unit tests
    - Pytests for the preprocessing pipeline currently include a dependency check, an environmental variable check, Azure upload and download verification, and a PDF conversion test. Each step of the preprocessing pipeline will eventually have a corresponding Pytest to ensure users can identify whether a specific portion is functioning correctly.
  - Expand, refine, run, and analyze evaluations
- **Dec 6th - Dec 11th:** Draft the final presentation and project write-up.
- **Dec 9th - Dec 11th:** Plan, coordinate, and rehearse presentation.

## Presentation Planning for December 12

- **Duration:** 20 minutes total
  - Introduction & Use Case: 5 minutes (Presenter 1)
  - Technical Implementation: 7 minutes
    - Presenter 2: Preprocessing Pipeline
    - Presenter 3: Chat Application
  - Testing and Results: 5 minutes (Presenter 4)
  - Q&A: 3 minutes

## Strategies for Planning Remaining Work

- Conduct weekly team syncs to monitor progress.
- Coordinate continuously on Slack
- Assign specific tasks to each team member to ensure equitable workload.

Narrow screenshot chat app:



The provided context does not include any discussions about property tax or any resident comments mentioning health issues or financial hardships during such discussions. The available information only covers issues related to noise and environmental impacts due to a rezoning request. If there are specific details or another document you would like me to search through, please provide that for further assistance.

Summarize my last meeting  
with the marketing team

Retrieve the project proposal  
document from the cloud

Find all files related to Q3  
sales reports

Check my upcoming  
meetings for today

Summarize the 'Budget  
Planning' document

List all files shared with me in  
the last week



Who gave the public comment relating to noise & environmental impact due to the rezoning request



The public comments relating to noise and environmental impact due to the rezoning request were given by Danny Caldwell, who expressed concerns about the noise generated while working on cars, and Mrs. Harrison, who expressed concerns about the environmental impact on her property and mentioned worries about noise.

Type your prompt



Evaluations:

Test case	Metric	Score	Status	Overall Success Rate
test_cases_2	Answer Relevancy	1.0 (threshold=0.8, evaluation model=gpt-4o, reason=The score is 1.00 because the response perfectly addressed the request without any irrelevant information. Great job!, error=None)	PASSED	100.0%
	Faithfulness	0.62 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 0.62 because the actual output contains several contradictions: it incorrectly states the Board of Commissioners Meeting date as February 20, 2024, instead of the correct date, January 2, 2024; it falsely claims there was a public comment section, which did not occur according to the retrieval context; and it inaccurately says no specific vote details were provided, despite the retrieval context mentioning 5-0 votes., error=None)	PASSED	
	Toxicity	0.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 0.00 because the actual output is entirely non-toxic, showing no signs of harmful language or negative connotations. Well done on maintaining a completely positive and respectful tone throughout., error=None)	PASSED	
test_cases_2	Contextual Precision	1.0 (threshold=0.7, evaluation model=gpt-4o, reason=The score is 1.00 because the relevant node in retrieval context, ranked first, effectively provides all the necessary details such as agenda items, motions, and votes for summarizing the Board of Commissioners Meeting. Great job on achieving perfect alignment!, error=None)	PASSED	100.0%
test_cases_2	Answer Relevancy	1.0 (threshold=0.8, evaluation model=gpt-4o, reason=The score is 1.00 because the response was perfectly relevant and directly addressed the inquiry without any irrelevant information. Great job!, error=None)	PASSED	100.0%
	Faithfulness	1.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 1.00 because there are no contradictions, indicating that the actual output perfectly aligns with the retrieval context. Great job maintaining accuracy!, error=None)	PASSED	
	Toxicity	0.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 0.00 because the output is completely non-toxic, showing no signs of harmful or offensive language. It likely uses positive and respectful language throughout., error=None)	PASSED	
test_cases_2	Contextual Precision	1.0 (threshold=0.7, evaluation model=gpt-4o, reason=The score is 1.00 because the relevant node in retrieval context perfectly describes the events of the meeting on March 19, 2024, with a detailed account of the presentation, public comments, and unanimous approval of TAP2-02 Mobile Food Vendors. Great job on the precision!, error=None)	PASSED	100.0%
test_cases_2	Answer Relevancy	1.0 (threshold=0.8, evaluation model=gpt-4o, reason=The score is 1.00 because the response is perfectly relevant and contains no irrelevant statements. Great job maintaining focus!, error=None)	PASSED	100.0%
	Faithfulness	1.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 1.00 because there are no contradictions, indicating perfect alignment between the actual output and the retrieval context. Great job!, error=None)	PASSED	
	Toxicity	0.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 0.00 because the output is entirely non-toxic and displays a respectful and positive tone throughout. It effectively avoids any language that could be considered harmful or offensive., error=None)	PASSED	
test_cases_2	Contextual Precision	1.0 (threshold=0.7, evaluation model=gpt-4o, reason=The score is 1.00 because the retrieval context comprehensively covers all the updates mentioned in the expected output for both meetings. Great job on providing all the necessary details!, error=None)	PASSED	100.0%
test_cases_2	Answer Relevancy	1.0 (threshold=0.8, evaluation model=gpt-4o, reason=The score is 1.00 because the response perfectly addresses the query without any irrelevant information. Great job!, error=None)	PASSED	100.0%
	Faithfulness	1.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 1.00 because there are no contradictions. The actual output aligns perfectly with the	PASSED	

	Toxicity	retrieval context., great job!, error=None) 0.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 0.00 because the actual output is completely non-toxic, demonstrating respectful and positive language throughout., error=None)	PASSED	
test_cases_2	Contextual Precision	1.0 (threshold=0.7, evaluation model=gpt-4o, reason=The score is 1.00 because the relevant node in the retrieval context perfectly addresses the input question with precise information, ensuring top-notch contextual precision. Great job!, error=None)	PASSED	100.0%
test_cases_2	Answer Relevancy	0.0 (threshold=0.8, evaluation model=gpt-4o, reason=The score is 0.00 because the output did not provide any relevant information about the top issues raised by residents, failing to address the input question., error=None)	FAILED	33.33%
	Faithfulness	0.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 0.00 because the actual output incorrectly claims there were no public comments made by residents between January 2024 and June 2024, while the retrieval context includes public comments made by Ali Bahr on January 16, 2024, and Jason McAlliffie on February 20, 2024., error=None)	FAILED	
	Toxicity	0.0 (threshold=0.5, evaluation model=gpt-4o, reason=The score is 0.00 because the actual output contains no toxic elements and is completely non-toxic. Great job on maintaining a positive and respectful tone!, error=None)	PASSED	