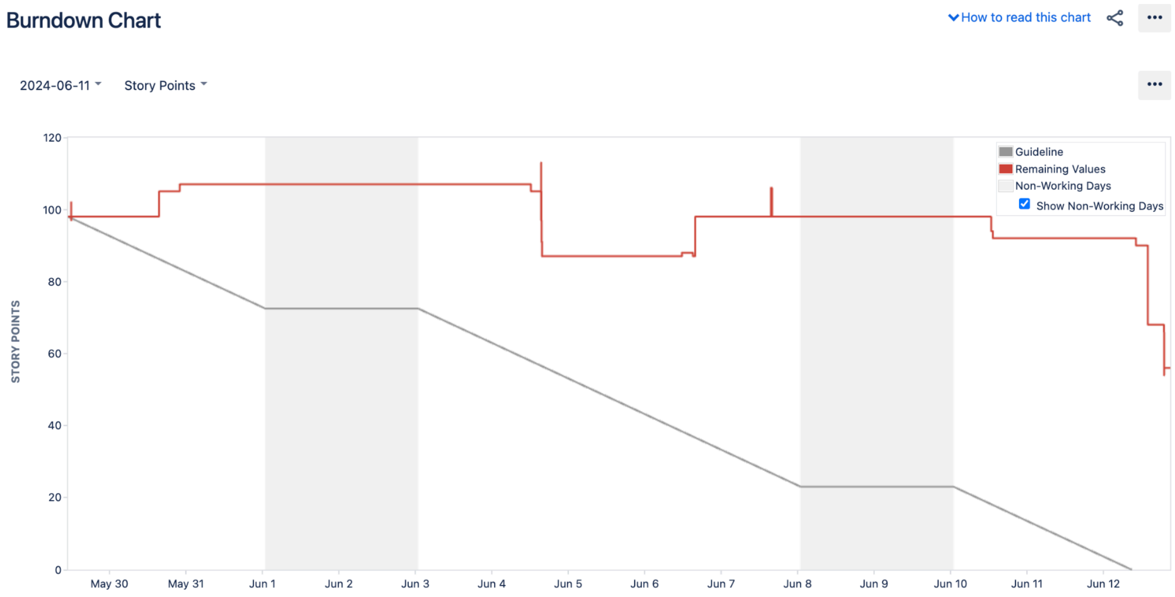
06.11 Sprint Summary & Review

Quick Links

📈 Sprint Performance

🔥 Sprint Burndown Chart



45% Complete | 31% Roadblocked |15% Incomplete 6% Rejected

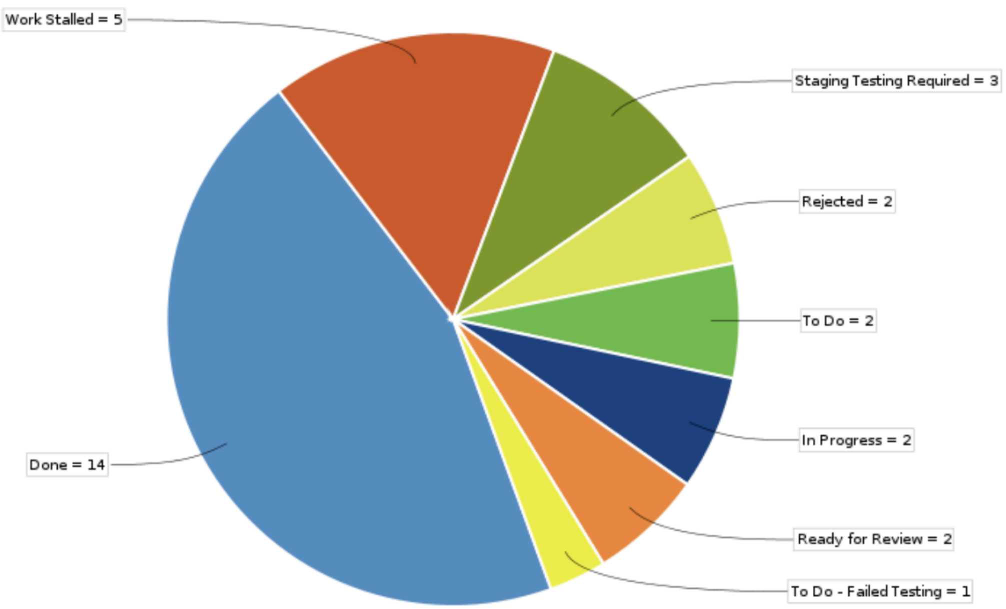
🚦 Issue Status Breakdown

Status	Issue Count	Percentage
Done	14	45%
Work Stalled	5	16%
Ready for Review	2	6%
Staging Testing Required	3	9%
In Progress	2	6%
To Do / Failed Testing	3	9%
Rejected	2	6%

LegalServer collaborate MainEvent

Filter: AIML 2024-06-11 (Status)

Chart



🫠 Issues that didn’t make the finish line

Incomplete

AIML-139 LS: Fake Criminal Charges Knowledge Transfer and Examples

- Building a proof of concept that demonstrates how AI can be leveraged to generate realistic looking but ultimately fake misdemeanor criminal rap sheets
- Greater specificity in the format of the document was required
    - A list of questions has been provided to LegalServer to confirm which fields should be altered and which should remain

Rejected

AIML-147 Investigate why Google API won't serve all networkninja users

- No longer an issue

Re: PolicyBot v.1.1.0 Release

Incomplete

AIML-131 Update base policy documents of model with the latest

- Updating PolicyBot to access the latest version of company policy documents

AIML-128 Policybot: Updates for Sentry

- Need to continue enhancing PolicyBot's Sentry error logging to capture and monitor cases where users are receiving "Something went wrong" messages
- Ongoing testing will ensure improved coverage and effectiveness of error logs

Work Stalled

AIML-132 Make policy docs automatically update/easily

- Implementing subroutines to automatically rebuild AI model based on triggers: i) updates to policy documents and ii) changes to the assistant models
- ⛔Blocked by AIML-204 Store policy documents in s3 Bucket; queued for next sprint

AIML-172 Clean up Branches in Policybot Repo

- Removing old and unnecessary branches from the PolicyBot repository, making it easier to manage and navigate
- The last remaining branch will be updated by AIML-130 Switch to slower Azure

Testing Blocked:

AIML-164 Move PolicyBot to GPT4o

- Updated PolicyBot's AI model to chatGPT 4o to improve its ability to return relevant and informed responses about company policies

AIML-130 Switch to slower Azure

- Moved PolicyBot to Azure API to improve security
- Not yet merged to staging; requires completion of AIML-132 Make policy docs automatically update/easily

Re: Sumit v1.2.0 Release

Incomplete

AIML-155 Draft Communication wrt Sumit improvements

- In preparation for the Sumit v1.2.0 release, we are
  - Developing content outlining the key updates and enhancements in the new version
  - Detailing explanations on how to effectively utilize the new features
  - Producing video demonstrations showcasing the updated workflows and navigation.

Work Stalled:

AIML-150 Sumit needs Google Workspace integration

- Upgrading Sumit to be integrated with NetworkNinja Google Workspace for a more streamlined user experience
- Sumit's Lambda server requires an API Gateway to allow communication between Google App Scripts and the Sumit application

Testing Blocked:

AIML-149 Refactor and Coordinate Sumit to function through Apps Script with Calendar information

- Updating Sumit to generate meeting summaries directly from users' Google Calendar events
- Testing on hold until an API Gateway that allows communication between Google App Scripts and Sumit application is established

AIML-161 Move Sumit to GPT 4o

- Upgraded Sumit to the latest GPT model, 4o
- Added an environment variable to facilitate easier updates in future
- Testing on hold until new user flow is in operation on stage

Re: LibreChat v0.7.2

AIML-167 Implement Sentry Error Logging LibreChat

- Attempted to implement Sentry error logging but discovered that Librechat does not currently support it
- Reached out via their Discord channel to request Sentry error logging support be included in a future release

Re: MainEvent Vision v1.3.0

Incomplete

AIML-122 Create the production version of the MainEvent Vision Lambda environment; AIML-120 Configure the API Gateway and corresponding API keys to interact with the FastAPI lambda functions

- Both issues are dependent on a functioning API Gateway, which unfortunately hit a roadblock due to a request size limitation

Work Stalled

AIML-125 Shut down the EC2 server currently running MainEvent Vision

- The existing set-up cannot be shut down until the new set up is operational

Rejected

AIML-123 Create the production version of the API gateway

- The API Gateway set up cannot handle requests larger than 10MB; unfortunately, this roadblocks all issues on the sprint relating to the migration of MainEvent Vision to a new infrastructure

📋 06.11 Summary (Brief)

In brief, the AIML team:

- 🧑‍🏫Hosted a workshop educating NetworkNinja employees on how to leverage the internal chatbot in their day-to-day work
- 📊 Continued to define metrics of interest that serve to measure the impact and success of our AIML tools
- 🧹Executed a comprehensive code pipeline and repository clean-up for our AIML tools
- 🐛 Started to integrate Sentry error logging across all AIML tools for improved bug monitoring
- 🔼 Began upgrading the suite of AIML tools to the latest GPT model
- 🔊 Built a proof of concept for a text-to-speech AI solution that would serve to streamline the creation of Collaborate's training videos
- 🔍 Developed a Python script to extract details from over 3,000+ Google Group threads
- 💡 Provided multiple solution paths on how AI can be leveraged for big data insights

🏃💨 What’s Next

Currently queued for the 06-25 sprint:

- Write the code to now *retrieve* the metrics of interest defined in earlier sprints and build a comprehensive dashboard
- Implement an s3 bucket and finalize the subroutine that auto-updates company policies for PolicyBot
- Release PolicyBot v1.1.0 with the latest ChatGPT model (4.0)
- Apply a number of improvements to LibreChat, including
  - Enabling streaming mode for bedrock models
  - Allowing our chatGPT login flow to allow access to *both* @networkninja.com and @legalserver.org email domains
- Continue optimizing our deployment pipelines for efficiency and reliability
- Conduct a meeting to discuss standardized management of secrets for AIML's AWS Lambda functions
- Apply a new strategy for moving MainEvent Vision to the updated infrastructure
- Release Sumit v1.2.0, showcasing the Google Workspace integration and improved user flow
- Develop a proof of concept for an AI assistant powered by forum answers and the LegalServer help site
- Create a build vs buy analysis for a text-to-speech AI tool, detailing the scope of work required for in-house development

📖06.11 Summary (Detailed)

🧑‍🏫Hosted a workshop educating NetworkNinja employees on how to leverage the internal chatbot in their day-to-day work

AIML-194 Host 'Mastering the Art of the Prompt' workshop

- Educated NetworkNinja employees on leveraging the internal chatbot, covering access, model selection, and *the art* of crafting effective prompts
- Post-workshop, shared a detailed slide deck and recording for company reference
- Increased employee awareness and proficiency in using the internal chatbot, enhancing overall utilization and effectiveness.

📊 Continued to define metrics of interest that serve to measure the impact and success of our AIML tools

AIML-105 Investigate & define available API Usage Metrics

- Selected Azure API key metrics of interest:
  - Total number of calls made to the Azure OpenAI API
  - Total number of generated tokens (output) from an Azure OpenAI model
  - Total number of inference tokens processed by an Azure OpenAI model
  - Recommended latency
- Selected AWS Bedrock key metrics of interest:
  - Invocations
  - Input/output tokens by model
  - Invocation client and server errors
  - Invocation latency

AIML-111 Identify associated Librechat usage costs, if any, to retrieve and/or store this data

- Provided breakdown of Amazon Timestream pricing
- Established a baseline understanding of costs
  - Exact calculations will follow once reports are running internally

🧹 Executed a comprehensive code pipeline and repository clean-up for our AIML tools

AIML-189 Add MODEL_SELECTION environment variable to Sumit stage and production

- Added the MODEL_SELECTION environment variable to enable easier updates, without requiring code changes, to the Sumit application

AIML-173 Update PolicyBot branch structure in Code Deploy Pipeline

- Implemented a new branch structure and more organized deployment process for PolicyBot

AIML-170 Update Code Pipeline for Sumit to use new branch structure

- Improved the deployment process by aligning it with the new branch structure for better organization and reliability.

AIML-159 Code Cleanup: Sumit-Lambda Branches

- Reviewed branches in the networkninja/sumit-lambda repository to identify those not related to any deployed infrastructure
- Reviewed the necessity of each branch and deleted unnecessary branches to streamline the repository and improve manageability.

🐛 Started to integrate Sentry error logging across all AIML tools for improved bug monitoring

AIML-171 Metric Investigation

- Investigated methods to measure OpenAI downtime that may have been impacting PolicyBot production environment performance
- Documented a method to implement downtime measurement but found the cost implications to be too high due to the need to monitor every component and check if all lambda functions are running
- Concluded that implementing this solution is not feasible from a cost perspective

AIML-168 Implement Sentry Error Logging for PolicyBot;

- Set up a Sentry error logging system for the PolicyBot application

AIML-165 Implement Sentry Error Logging for MainEvent Vision

- Set up a Sentry error logging system for the MainEvent Vision application

🔼 Began upgrading the suite of AIML tools to the latest GPT model

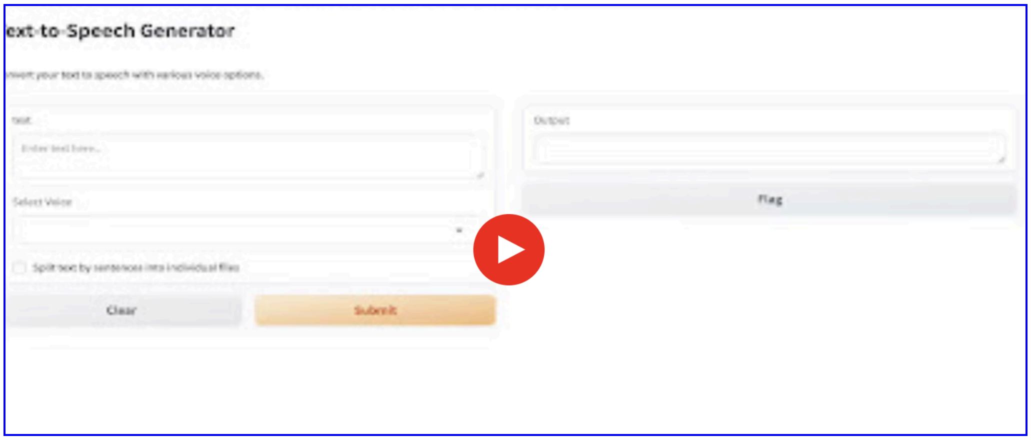
AIML-162 Move MainEvent Vision to GPT4o

- Upgraded MainEvent Vision to latest GPT model, 4o

🔊 Built a proof of concept for a text-to-speech AI solution that would serve to streamline the creation of Collaborate's training videos

AIML-144 Deliver barebones proof of concept (text-to-speech feature)

- Used the text-to-speech API to generate an audio file from a sample paragraph
- Created a proof of concept demonstrating how the audio file can be segregated into multiple audio files and how a user can select from a catalogue of different AI-generated voices



ext-to-Speech Generator

nvert your text to speech with various voice options.

⏯️[WATCH VIDEO] Click above image to watch Text-to-Speech Audio File Generator

Note: Video is accessible by all users with a @networkninja.com email domain

🔍 Developed a Python script to extract details from over 3,000+ Google Group threads

AIML-138 LS: Demonstrate scraping the ls siteadmins list for useful information for an assistant

- Gathered all 3,000+ Google Group threads from the LegalServer siteadmin list using a Python script to automate the web scraping process
- Developed an additional script to extract and store sender's name, submission date, email content (raw HTML) and link to the original thread in JSON
- This work serves to unblock the creation of an AI assistant that mines this database to autogenerate answers to commonly asked questions

💡 Provided multiple solution paths on how AI can be leveraged for big data insights

AIML-41 Research and Questions: Internal Jira ticket duplicate analysis

- Investigated ways in which AI can be leveraged to analyze spreadsheets or databases and return insights or rule-based alerts
- Provided a number of possible solution paths, including using NNI's internal chatbot for data uploads and queries
- Developed a concept of using SQL generation and queries for data analysis
  - Created a custom SQL assistant (*SQLBot* 🤖) to generate SQL commands for specifics queries
- Reviewing scope to determine if proof of concept will be necessary or if a *Saas* product will be sufficient