Agentic AI with Orchestrated Workflows for Engineers

Clair J. Sullivan
clair@clairsullivan.com

CLAIR SULLIVAN
& ASSOCIATES

bill

Schedule

9:00 - 9:30	Module 0: Introduction to Agentic AI
9:30 - 11:30	Module 1: Foundational Agentic Patterns with n8n
11:30 - 2:00	Module 2: Advanced Patterns & Custom API Integration
2:00 - 4:00	Module 3: LLM Evaluation with Braintrust
4:00 - 4:30	Awards!

Introduction

Agenda

- Brief introduction to the people in the room
- Structure for today
  - *Brief* lectures
  - Activities
  - Challenges (what is scored!)
- Module 0: Introduction to Agentic AI
- Module 1: Foundational Agentic Patterns with n8n
- Module 2: Advanced Patterns & Custom API Integration
- Module 3: LLM Evaluation with Braintrust
- Wrap up, announcement of winners!

What We CAN Do Today

- Introduce all of the basic concepts that you need in order to start creating GenAI applications in n8n and doing evaluations in Braintrust
- Keep it simple
- Basic workflows for GenAI
- Make some things that work!

What We Won’t Be Able to Cover Today

- Going beyond the basics
- Discussing every different approach to coding a problem
  - Includes all of the different ways to do a thing with n8n and its ever-evolving suite of integrations
- Use a ton of n8n integrations
  - Keeping it simple with Google
- Go in depth on LLM evaluations
- Use PII, real Bill data
- Discuss productionalizing GenAI, AIOps

A brief survey:

PollEv.com/clairsullivan399

Module 0: Introduction to Agentic AI

What is Generative AI (GenAI)?

- Traditional AI (AKA machine learning) focuses on using math to make predictions
- GenAI **creates** new content
- Trained on large datasets using deep learning to learn patterns and generate novel output
- Examples of GenAI in action

How Large Language Models (LLMs) Work

- Transformer models: the backbone of LLMs
  - Captures relationships between words across long texts
  - Key innovation in models like GPT, BERT, etc.
- Text is broken into **tokens**
- LLMs have a maximum number of tokens they can process: **context window**
  - When exceeded, older tokens are forgotten (“sliding window effect”)
  - Varies by model
  - Longer context ≠ perfect memory
- LLMs are asked to do things through their **prompts**
- Limitations of LLMs
  - **Hallucinations**
  - Training data limitations (date, subject matter)
  - Computation cost

Agents

- “An agent in generative AI is a system that uses a language model to make decisions, take actions, and interact with tools or environments based on a defined goal or task.”
- An LLM on its own can only use language to do things
- Agents extend this by giving the LLM **tools** for taking action
- Can plan and reason to decide on what tool to use
- Useful for multi-step tasks, orchestrating actions

Tools

- The true power behind agentic GenAI!
- External functions or APIs that LLMs can call to perform specific tasks
  - Ex: looking up information, doing math, querying a database, retrieving files from storage
- Extend LLM's capabilities beyond text generation, enabling it to interact with real-world systems, retrieve fresh data, etc.
- Tool descriptions should very clearly state what the tool does, what information it requires, etc.

Agents are Different Than Chains

- Chains: a pre-defined sequence of steps
  - The exact order or execution is known
  - Each step is connected to another step in a predetermined way
  - No branching
  - Can’t adapt
- Agents go beyond chains by:
  - Can make decisions about which actions to take based on intermediate results
  - Can loop and iterate
  - Can handle conditional paths
  - Can adjust approach based on results, detection of failures

You can create workflows with LLM chains in n8n. However, in this workshop we will focus on agentic workflows for the reasons above.

Retrieval Augmented Generation (RAG)

- LLMs are trained off of a large corpus of knowledge
  - They understand general concepts, NOT specifics
  - When they don’t know the specifics, they tend to hallucinate
- RAGs providing them with additional information on relevant subjects
- Involves giving the LLM access to external data sources to improve the accuracy and relevance of generated responses
  - Documents, databases, etc.
  - Retrieves relevant context at query time
- Essential for domain-specific tasks when it is unlikely that the relevant information is in the LLM’s training set

How Do We Do That?

Knowledge Base

Chunks

Database

Retrieved Chunks

Response

Text Embedding Model

Large Language Model

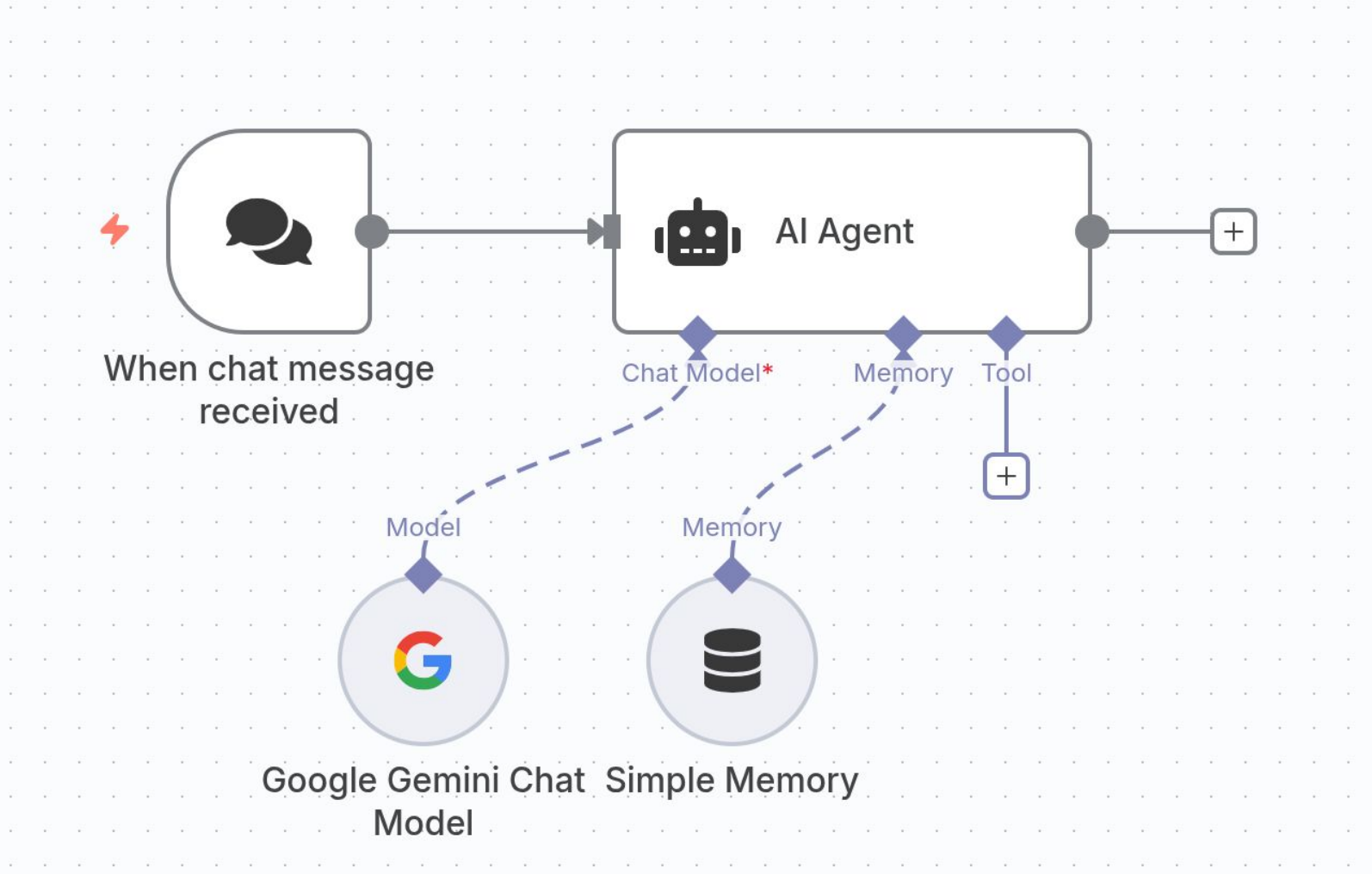
What are the Embeddings?

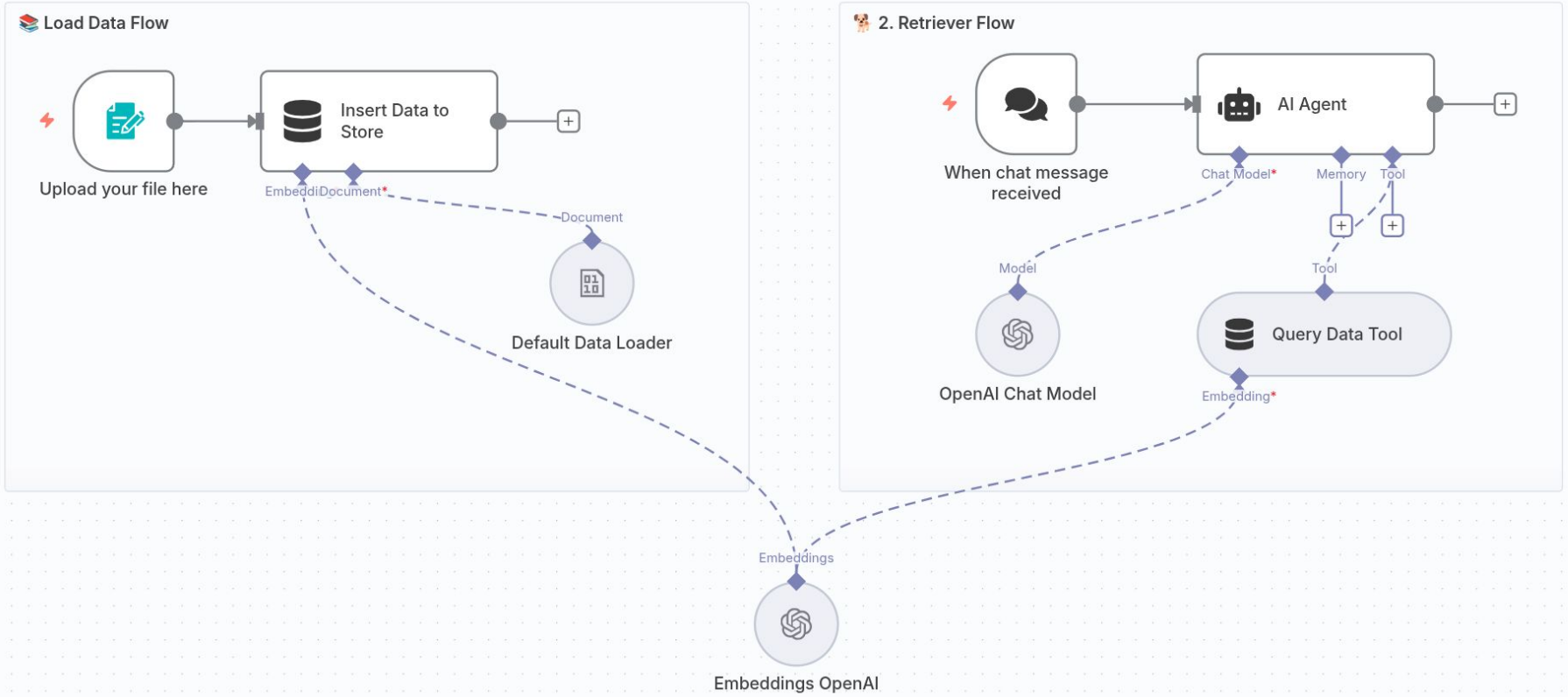
- All GenAI is just math
- Need to convert text into numbers so that the LLM can do math on them
  - Predictions of next most likely word given previous words
- An n-dimensional **vector** (a list of floats) with each value between -1.0 and 1.0
  - Which model you use determines n, but numbers > 1000 are typical
- Stored in a database for later retrieval based on **similarity**
  - Cosine similarity is easy to do on vectors, which allows the LLM to quickly find the most relevant pieces of information in your database
- Remember: everything is about converting strings (or other data types) to vectors, meaning you need to start with strings!

Module 1: Foundational Agentic Patterns with n8n

What is n8n?

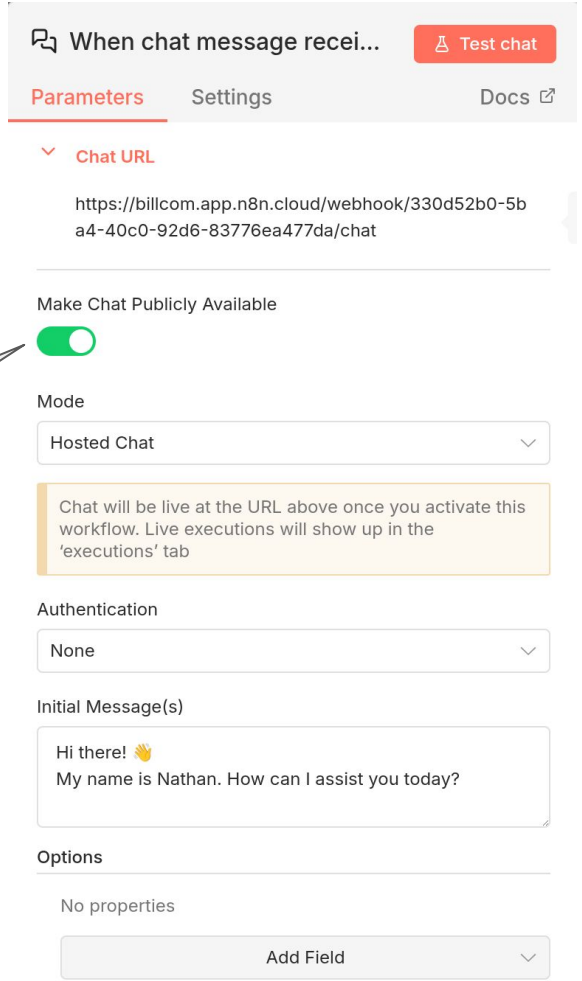
- Low-code workflow automation tool
- Enables agentic AI workflows by easily combining agents with integrations and services
- Workflows can interact with apps, execute tasks without constant human input
- Provides visual workflows to allow monitoring capabilities, logging
- Can run in hosted or self-hosted environments
- Strong user community with many free templates
- When basic integrations fall short, allows you to create custom nodes with JavaScript or Python


When chat message received
AI Agent
Chat Model*
Memory
Tool
Model
Memory
Google Gemini Chat Model
Simple Memory


Load Data Flow
Upload your file here
Insert Data to Store
EmbeddiDocument*
Document
Default Data Loader
2. Retriever Flow
When chat message received
AI Agent
Chat Model*
Memory
Tool
Model
OpenAI Chat Model
Tool
Query Data Tool
Embedding*
Embeddings
Embeddings OpenAI

Easily Make Chat Public

- Chat URL is a webhook under the hood
- Allows you to embed as a chat widget or on any webpage or app



80% of Workflows Use a Limited List of Nodes

- Logic nodes
  - IF
  - SWITCH
- Combining data sources
  - MERGE
  - AGGREGATE
- Connectivity
  - HTTP Request
  - Webhooks
- Code

Structure for the Hands-On Work

- Start with the **onboarding.md** document in Module 0 to get all accesses set up
- You will begin each module by going through the “Activity” documents to learn the skills needed to do the “Challenges”
- **Always do the Activities before the Challenges!**
- You will be scored based on the “Challenges”
- You don’t have to submit solutions for every question
- You are *strongly* encouraged to work as a team, plan out the work, “divide and conquer”

Ground Rules

- **Have fun!!!**
- Experiment
- Engage with everyone on your team
- Extra points can be awarded at any time for great *shared* learnings
- There are many prizes up for grabs, not just high scores
- *The Honor System: many of the challenges can be solved to some extent with traditional programming. For the answers you submit, please be honor-bound to submit your answers created via a GenAI solution.*
- *The Honor System II: Resist the urge to copy and paste your team's solutions. Try to create your own answer, which will likely be a different result than your teammates.*

https://github.com/ClairSullivan-Associates/n8n_agentic_course

https://aiworkshop.dev.bill.com/

Schedule

9:00 - 9:30	Module 0: Introduction to Agentic AI
9:30 - 11:30	Module 1: Foundational Agentic Patterns with n8n
11:30 - 2:00	Module 2: Advanced Patterns & Custom API Integration
2:00 - 4:00	Module 3: LLM Evaluation with Braintrust
4:00 - 4:30	Awards!