Virtual Assistant Agent for Policy Reasoning

Tianbing Xu

Agent

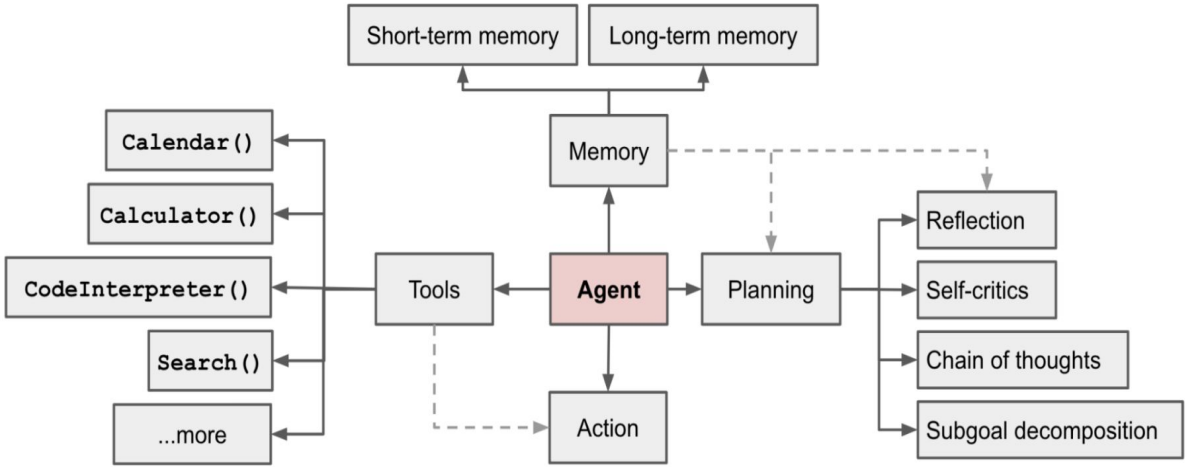


Fig. 1. Overview of a LLM-powered autonomous agent system.

System Design

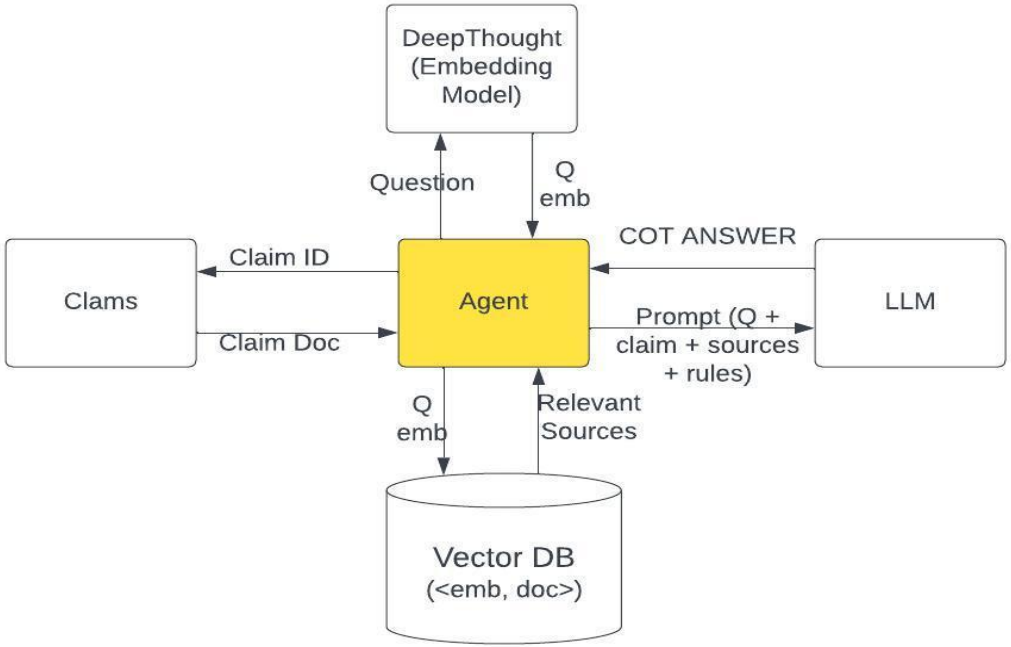
Availability, Performance, Scalability

Project criticality - Tier 2

SLOs outlined in this document aim for an overall target of 99.9% availability. Latency for dependent services may fluctuate per request type, reflecting the varying degrees of complexity.

Bighead DeepThought Embedding Serving: ~20 ms
Vector DB Retrieval : ~100 ms
OpenAI chat completion API (GPT-4) for QA with Chain-Of-Thought reasoning: average ~2.5 seconds (source)
Total Latency: average ~2.85 seconds (source)

System Architecture



Components & Interfaces

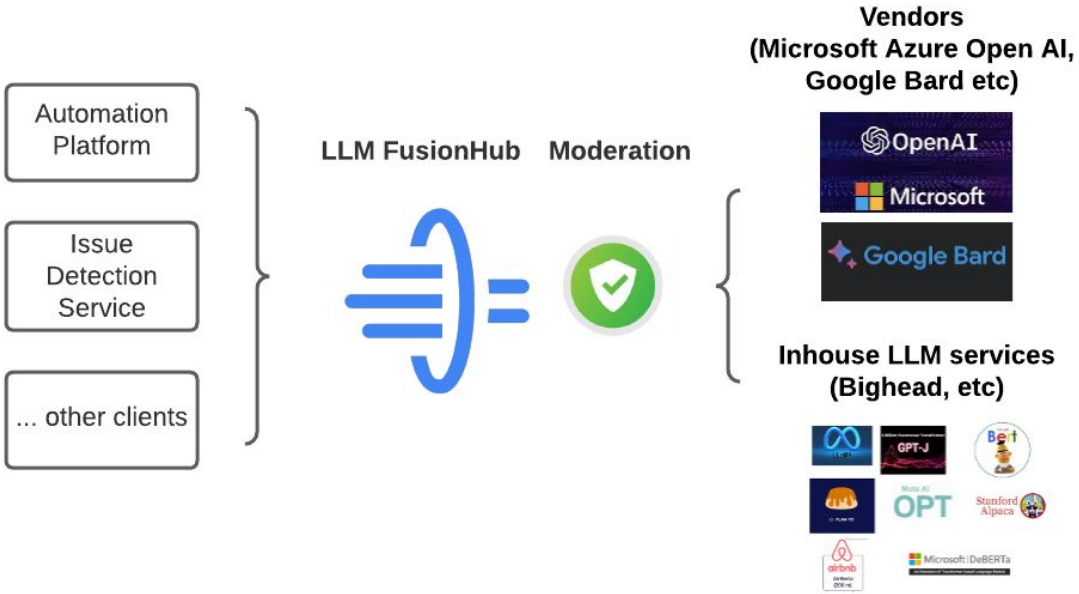
Embedding

Embedding vectors for questions and HDP Policy terms document chunkings. Sentence Transformer or General Text Embeddings (GTE) model

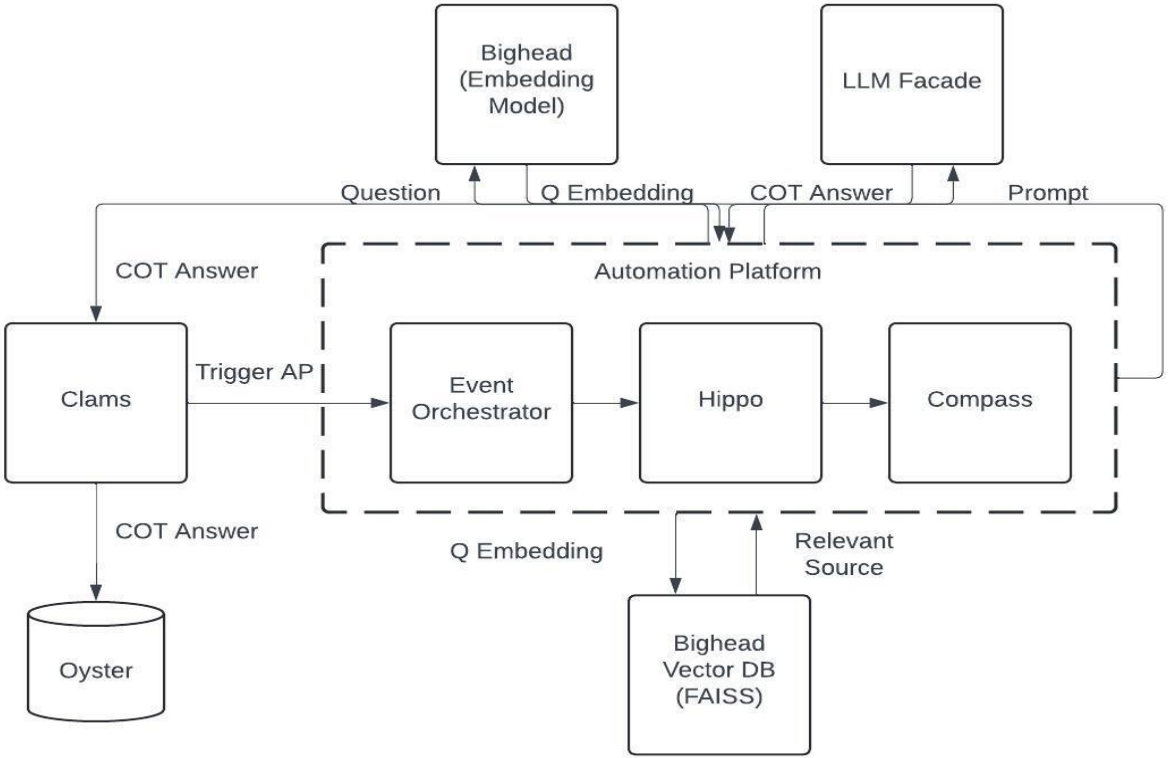
Vector DB

FAISS + DeepThought
FAISS (Facebook AI Similarity Search)
Indexed with the embedding of HDP Policy document chunks, allowing for deployment onto Bighead to facilitate relevant document retrieval.

LLMs

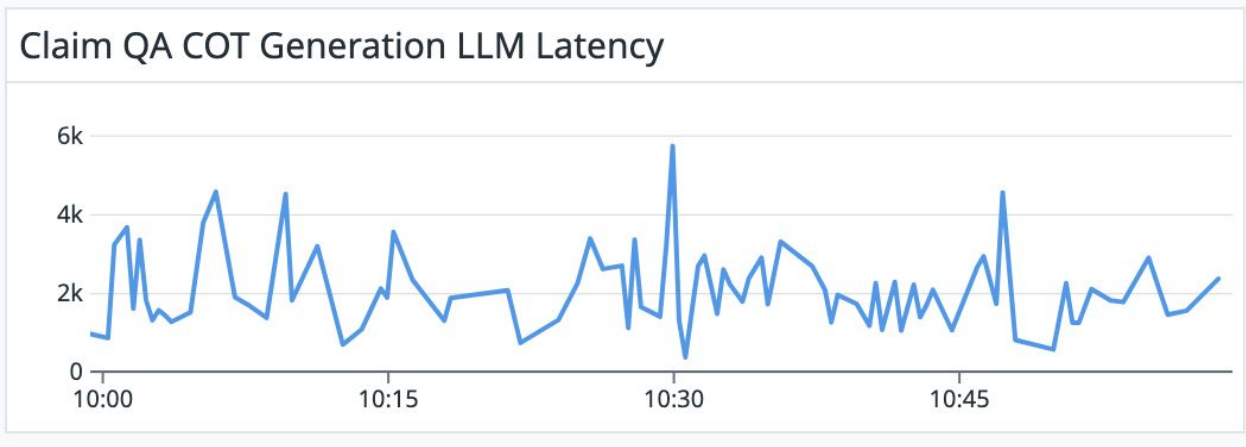

Automation Platform
Issue Detection Service
... other clients
LLM FusionHub
Moderation
Vendors (Microsoft Azure Open AI, Google Bard etc)
OpenAI
Microsoft
Google Bard
Inhouse LLM services (Bighead, etc)
GPT-J
Bert
OPT
Stanford Alpaca
airbnb
Microsoft DeBERTa

Agent System building with Automation Platform



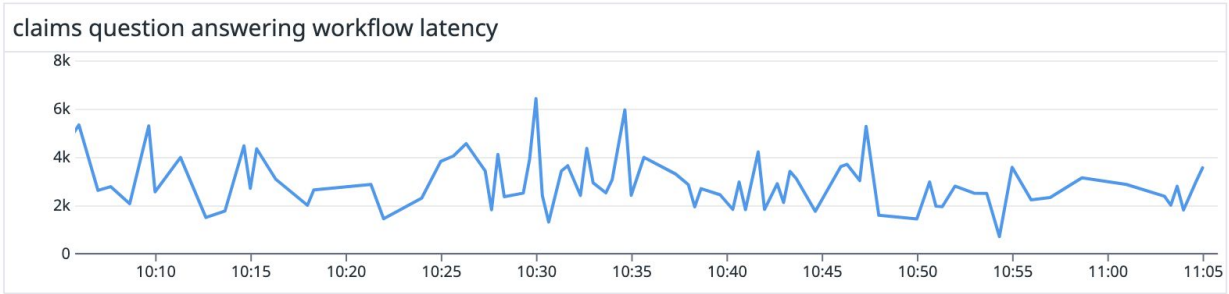
Result

Latency (LLM)


Claim QA COT Generation LLM Latency
6k
4k
2k
0
10:00
10:15
10:30
10:45

Latency (Agent Workflow)

AP: claimQAReasoningBot workflow



LLM Accuray and Cost to Serve

Model	Tokens per minute	**Estimated Cost Per Day**	**Accuracy**
gpt-4	60k	$100	0.80
gpt-4-32k	190	$300	0.85

Example in Productions

Claim item eligibility signals

Possible ineligible item(s) or issue detected
YES

Explanation:
Yes. The claim includes the situation of "Extra guests stayed without notifying me" which falls under the "Extra guest fee" and

reference_id	claim_amount_usd	claim_overview	claim_items	explanation	decision
CLSF-04350295	163.59	I'm awaiting dryer repair, the washer & dryer were used to wash/dry sandy items by you or your guests. The gravel is behind the drum & needs to be taken apart & cleaned out as this is a hazard. I've had it looked at & accessed however this is a job for a professional appliance repair company. I'm not able to use the dryer as I'm waiting on the appliance repair company to do the work that needs to be done and will submit a claim amount for that work. It was clearly stated in the house	["Broken Reclining wood handle beach chair", "Dryer drum has a lot of gravel in it"]	YES. The items not covered are the repair of the dryer and the broken reclining wood handle beach chair. HIGH	YES

Product Impacts

- 1.2% reduction in autopaid claims
- 46% of claim blocked from auto payment are declined in full
- 13% reduction in payout rate for paid claims

Reference

Lilian Weng, et al. LLM-powered Autonomous Agents, 2023

Jason Wei, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, NeurIPS 2022

Shunyu Yao, et al. ReAct: Synergizing Reasoning and Acting in Language Models, ICLR 2023