Allegro Pay Backstage Assitant

Owner: Paweł Piwowarczyk
Reviewer:
Contributors: adam.trepka@allegro.com, pawel.piwowarczyk@allegro.com
Date Generated: Mon Nov 04 2024

Executive Summary

High level system description

Analyze questions people ask to grab relevant key-words used to search the Technical Documentation index with search API delivered by Backstage. The most relevant answers are then passed to the context of the model which tries to find the best answer to the question.

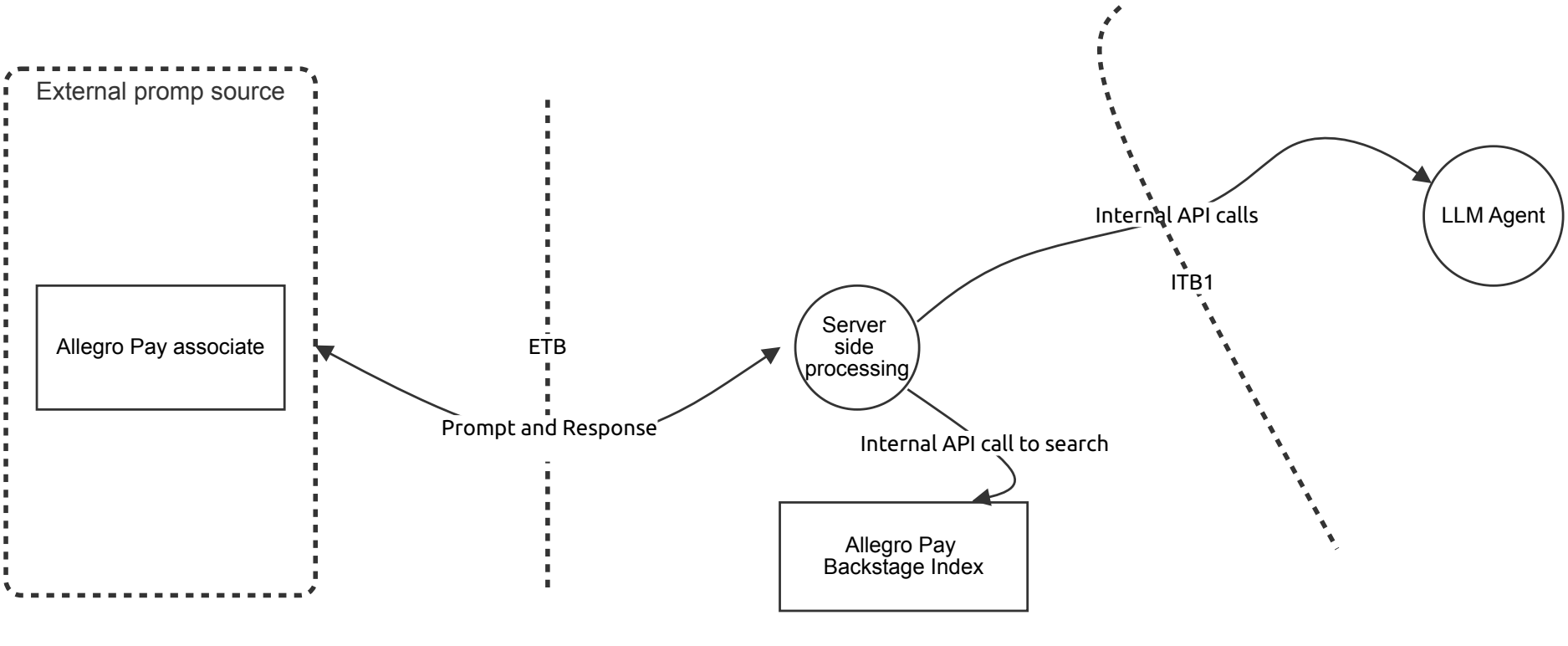
We want to solve problem: Huge amount of questions to Allegro Pay technical platform on help channels (at least 5-10 daily requests) that some part (approximately 50%) might be answered with Technical Documentation stored in Allegro Pay Backstage

Summary

Total Threats	6
Total Mitigated	6
Not Mitigated	0
Open / High Priority	0
Open / Medium Priority	0
Open / Low Priority	0
Open / Unknown Priority	0

DFD

DFD for Backstage Assitance



DFD

Allegro Pay associate (Actor)

Number	Title	Type	Priority	Status	Score	Description	Mitigations
4	V001: Prompt injection	Spoofing	Medium	Mitigated		Prompt injection - System prompt modification - User can generate malicious input, that will overwrite its controls (by the jailbreak)	The user's query is not sent directly to the search engine backstage. First, we detect the intent of the question and convert it into keywords, which are then passed to the search engine. We are able to precisely determine the scope of documents available to the user. This way, we protect the system from leaking sensitive content.

Allegro Pay

Backstage Index (Actor)

Tech Documentation, wiki, announcements and q&a available in Allegro Pay Backtage

Number	Title	Type	Priority	Status	Score	Description	Mitigations

LLM Agent (Process)

Number	Title	Type	Priority	Status	Score	Description	Mitigations
6	V003 User sensitive information disclosure to model vendor	Tampering	Medium	Mitigated		V005: Attack on vendor infrastructure resulting in model backdoor.	The assistant does not have defined functions that it could execute "autonomously." The responses are based solely on pre-selected documents indexed in the Backstage system. The assistant in the Backstage system is only accessible to users with the appropriate role/permissions. Access via Slack is limited to a strictly defined group of channels. It is not possible to contact the bot through a private message or add it to a channel without consulting the team responsible for the development and maintenance of the assistant. Access to the assistant's API is secured with a bearer token, preventing this type of attack from being automated. The assistant uses models available within Azure OpenAI Services. We are able to specify a particular version of the model. We do not use the "latest" or "preview" versions.
12	New STRIDE threat	Information disclosure	Medium	Mitigated		V006: LLM is unable to filter sensitive information (vendors researches in progress)	The Backstage system, which serves as the data source for the assistant, does not contain confidential information that Allegro Pay employees should not access. The assistant does not have defined functions that it could execute "autonomously." The responses are based solely on pre-selected documents indexed in the Backstage system.

Internal API call to search (Data Flow)

Number	Title	Type	Priority	Status	Score	Description	Mitigations

Internal API calls (Data Flow)

Number	Title	Type	Priority	Status	Score	Description	Mitigations

Prompt and Response (Data Flow)

Number	Title	Type	Priority	Status	Score	Description	Mitigations
9	New STRIDE threat	Tampering	Medium	Mitigated		V002: Modification of model parameters (temperature, p, model version)	The user's query is not sent directly to the search engine backstage. First, we detect the intent of the question and convert it into keywords, which are then passed to the search engine.
10	New STRIDE threat	Information disclosure	Medium	Mitigated		V003: User sensitive information disclosure to model vendor (user behavior - data in prompt)	The user's query is not sent directly to the search engine backstage. First, we detect the intent of the question and convert it into keywords, which are then passed to the search engine.

Server side processing (Process)

Lang Chain and other processing of the input

Number	Title	Type	Priority	Status	Score	Description	Mitigations
11	New STRIDE threat	Denial of service	Medium	Mitigated		V004: Lack of rate limiting can overflow the budget resulting in DoS from insufficient funding.	Access to the assistant's API is secured with a bearer token, preventing this type of attack from being automated. We have not implemented other security methods such as rate limiting.