

Requests



Alejandra Caggiano

Observabilidade



GeneXus Enterprise AI armazena e rastreia cada solicitação, fornecendo às organizações uma visibilidade completa do uso de seus assistentes, modelos de inteligência artificial e o custo associado a cada solicitação.

Observabilidade

- Monitorar e analisar o uso de recursos
- Tomar decisões informadas
- Otimizar o uso de recursos.
- Gerenciar fluxo de trabalho
- Controlar os gastos.
- Maximizar o retorno do investimento

Isto permite que as organizações monitorem e analisem o uso de recursos, tomem decisões informadas sobre a alocação desses recursos e otimizem o seu uso para alcançar rentabilidade.

Com uma visibilidade clara, tanto do uso quanto do custo, as organizações podem gerenciar de forma eficaz seus fluxos de trabalho impulsionados por inteligência artificial, controlar os gastos e maximizar o retorno do investimento.

Observabilidade

As Organizações podem manter o controle sobre sua infraestrutura de inteligência artificial,

Ao aproveitar estas características, as organizações podem manter o controle sobre sua infraestrutura de inteligência artificial, identificar áreas de melhoria e tomar decisões baseadas em dados para melhorar a eficiência operacional.

Requests

The image displays three overlapping screenshots of the GeneXus Enterprise AI interface, illustrating the 'Requests' monitoring functionality.

Top Screenshot: Request List

This view shows a table of requests with the following columns: Module, Assistant Name, AIModel Name, Api Token Name, Input, Timestamp, Time (Ms), and Status. The data is filtered by 'This week' and 'Module'. The table shows several successful requests from the 'MarketingAssistant' module.

Module	Assistant Name	AIModel Name	Api Token Name	Input	Timestamp	Time (Ms)	Status
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:52	832	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:42	796	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:40	804	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:40	796	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:40	257	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:40	509	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:39	432	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:38	827	Succeeded
MarketingAssistant	MarketingAssistant	gpt-3.5-turbo-03b	Default	chat	14/10/2024 10:55:37	58	Succeeded
MarketingAssistant	N/D	N/D	Default		14/10/2024 10:55:37	37	Succeeded
MarketingAssistant	N/D	N/D	Default		14/10/2024 10:55:36	40	Succeeded
MarketingAssistant	N/D	N/D	Default		14/10/2024 10:55:35	46	Succeeded

Middle Screenshot: HttpProxy Request Log

This view shows the detailed log for a specific request. It includes fields for Log ID, Log Name, Log Date, and Log Timestamp. The log content shows the request body, including headers and the chat input.

```

Log ID: 1
Log Name: http://chat
Log Date: 14/10/2024 10:55:52
Log Timestamp: 14/10/2024 10:55:52

{"messages": [{"role": "user", "content": "chat"}]}
  
```

Right Screenshot: HttpProxy General (Pretty)

This view shows the general information for a request, including the timestamp, the assistant name, and the input text.

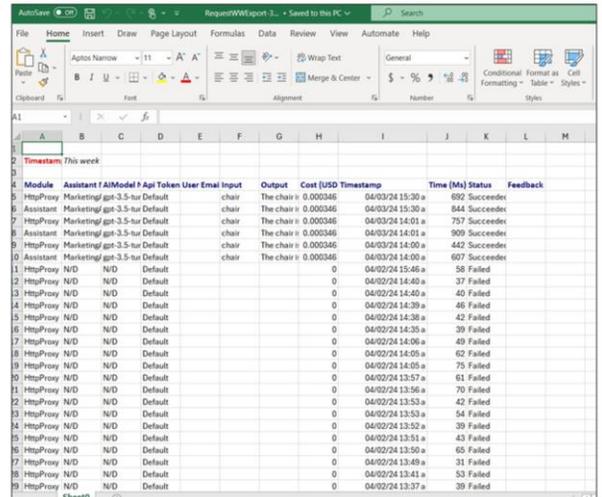
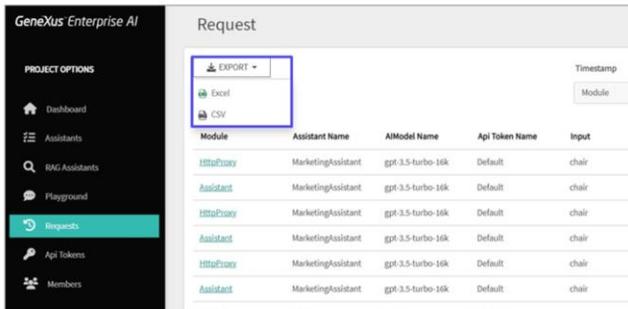
```

General (Pretty)
Timestamp: 14/10/2024 10:55:52
Assistant Name: MarketingAssistant
Input: chat
Output: The chat is a text with a back, usually with few logs, and that can only fit one person.
  
```

A partir da plataforma, no menu esquerdo, a opção “Requests” dá acesso a um monitoramento integral que fornece observabilidade completa de cada solicitação realizada.

Esse monitoramento permite filtrar facilmente solicitações por modelo, assistente, por api token, intervalo de data e hora e status, o que permite, por sua vez, identificar rapidamente solicitações de interesse específicas. Além disso, ao clicar na coluna Module de uma solicitação em particular, podemos acessar sua informação detalhada.

Requests



Dentro desses detalhes podemos ver os dados de entrada e saída, o modelo específico utilizado para a solicitação, o custo associado e o carimbo data/hora que indica quando foi executada a solicitação.

Essa capacidade de acessar e revisar os detalhes completos de cada solicitação permite compreender os dados e processos subjacentes, o que, por sua vez, facilita poder identificar e abordar qualquer problema ou área de melhoria.

Se voltarmos ao painel inicial, vemos que também é possível exportar este detalhe, por exemplo, para Excel.

Com esta característica de observabilidade integral, as organizações podem garantir a precisão, eficiência e rentabilidade de seus fluxos de trabalho impulsionados por Inteligência Artificial em GeneXus Enterprise AI.

GeneXus[™]
by **Globant**

training.genexus.com