



# A.I. DAY 2025 RELOADED

Come utilizzare le nuove funzionalità AI  
in SQL Server 2025 con Ollama



Danilo Dominici





---

## Platinum Sponsor

---



---

## Gold Sponsor

---



---

## Technical Sponsor

---



# Agenda

Introduzione alle novità AI in SQL Server 2025

Cos'è Ollama e come si integra con SQL Server

Best practices e sicurezza

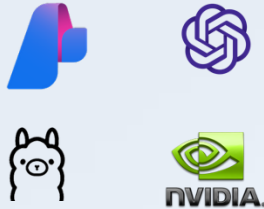


# Introduzione alle novità AI in SQL Server 2025



# SQL Server 2025

The AI-ready enterprise database from ground to cloud



**AI built-in**



**Develop modern  
data applications**



**Integrate your data  
with Fabric**



**Secure by default**



**Mission critical engine**



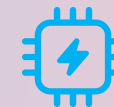
**Connected  
with Arc**



**Assisted by  
Copilots**



**Benchmark  
leader**



**Optimized for  
latest hardware**




**Built for  
all  
platforms**

**Version**

SQL Server 2025 Preview ▾

 Filter by title

- SQL Server
- Docs navigation tips
- Previous versions 2005-2014
- ▾ Overview
- What is SQL Server?
- Connect to the Database Engine
- ▾ What's new?
- SQL Server 2025 Preview
- SQL Server 2022
- SQL Server 2019
- SQL Server 2017
- SQL Server 2016
- > Editions and features
- > Release notes
- > Known issues
- > AI
- > Business continuity
- > Database design
- > Development
- > Internals & architecture
- > Installation
- > Migrate & load data
-  Download PDF

[Learn](#) / [SQL](#) / [SQL Server](#) /[Ask Learn](#)[Focus mode](#)

⋮

# What's new in SQL Server 2025 Preview

10/06/2025

**Applies to:**  SQL Server 2025 (17.x) Preview

SQL Server 2025 (17.x) Preview builds on previous releases to grow SQL Server as a platform that gives you choices of development languages, data types, on-premises or cloud environments, and operating systems.

This article summarizes the new features and enhancements for SQL Server 2025 (17.x) Preview.

**In this article**

- Get SQL Server 2025
- Release candidate 1
- Feature highlights
- AI
- Developer
- New developer editions
- Analytics
- Availability and disaster recovery
- Security

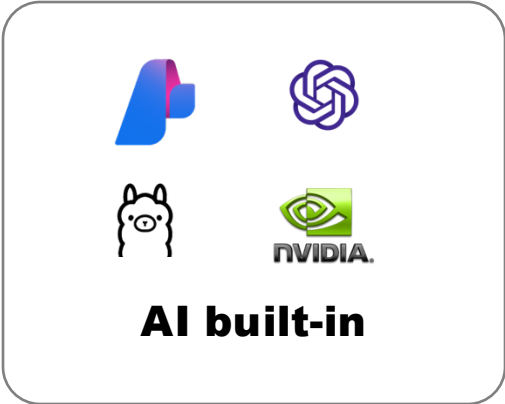
[Show 15 more](#)**Was this page helpful?**

 Yes

 No



# SQL Server 2025: AI



<b>Vector data type</b>	Store vector data optimized for operations such as similarity search and machine learning applications. Vectors are stored in an optimized binary format but are exposed as JSON arrays for convenience. Each element of the vector can be stored either using a single-precision (4-byte) or half-precision (2-byte) floating-point value.
<b>Vector functions</b>	New scalar functions perform operations on vectors in binary format, allowing applications to store and manipulate vectors in the SQL Database Engine.
<b>Vector index</b>	<p>Create and manage approximate vector indexes to quickly and efficiently find similar vectors to a given reference vector.</p> <p>Query vector indexes from <code>sys.vector_indexes</code>. Requires <code>PREVIEW_FEATURES</code> database scoped configuration.</p>
<b>Manage external AI models</b>	Manage external AI model objects for embedding tasks (creating vector arrays) accessing REST AI inference endpoints.





# Quali problemi cerchiamo di risolvere con l'AI ?

- Ricerche più “intelligenti” 😊 sui nostri dati
- Inglobare anche altri documenti per centralizzare la ricerca
- Creare *building blocks* riutilizzabili – assistenti intelligenti, RAG, Agenti AI
- Fruire dell'AI in modo sicuro e scalabile
- Usare il linguaggio T-SQL a noi familiare anche per attività complesse come quelle che coinvolgono l'AI



# Costruire applicazioni enterprise AI-ready

Build Agentic RAG patterns  
*Inside the engine*



## Vector Store

Native vector data type and DiskANN index



## Model Management

Declarative Model definitions ground/cloud



## Embeddings built-in

Text Chunking and built-in multimodal embedding generation



## Simple semantic searching

Vector distance (KNN) and Vector search (ANN)



## Framework integration

LangChain, Semantic Kernel, EF Core

# Sicurezza, SQL, e l'AI



Controllo degli accessi tramite la sicurezza di SQL Server



Controllo sul modello da utilizzare



Modelli locali o cloud isolati da SQL Server



Utilizzo di RLS, TDE, e Dynamic Data Masking



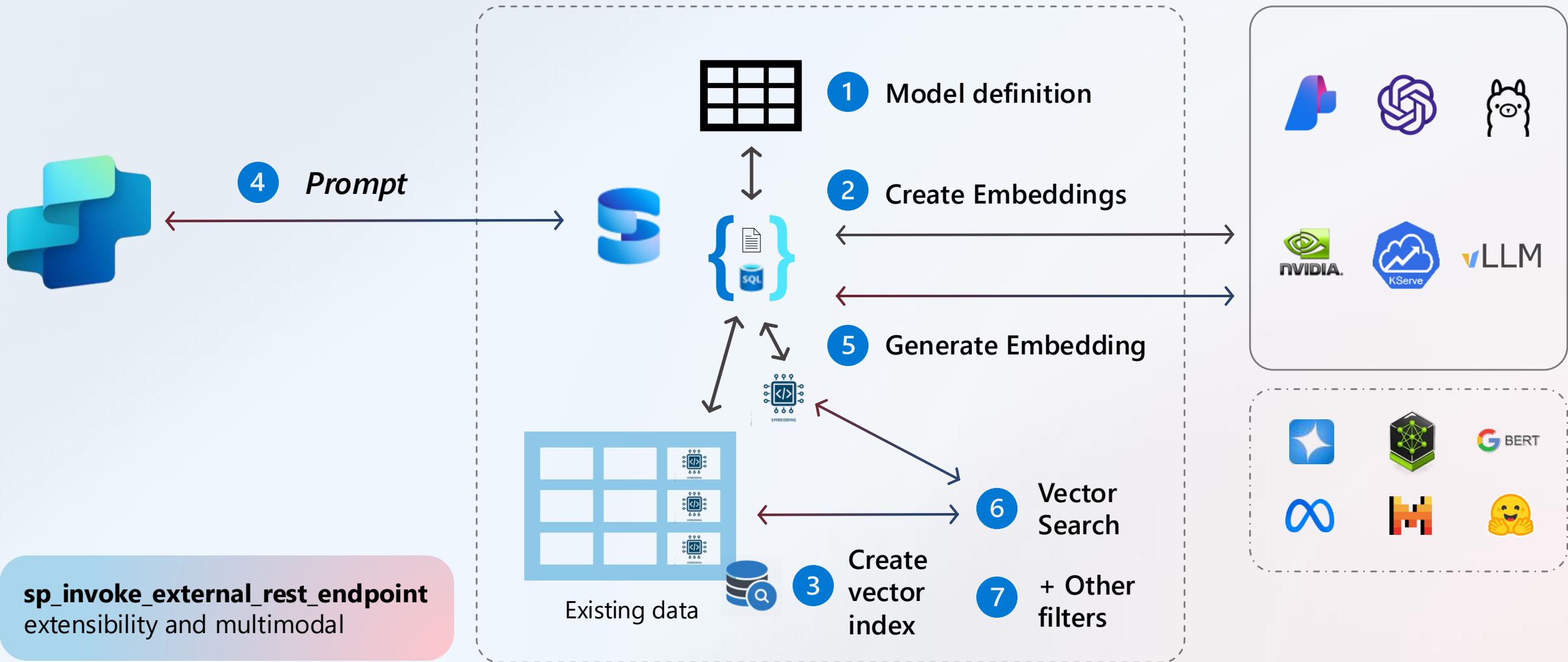
Tracciabilità degli eventi con l'auditing di SQL Server



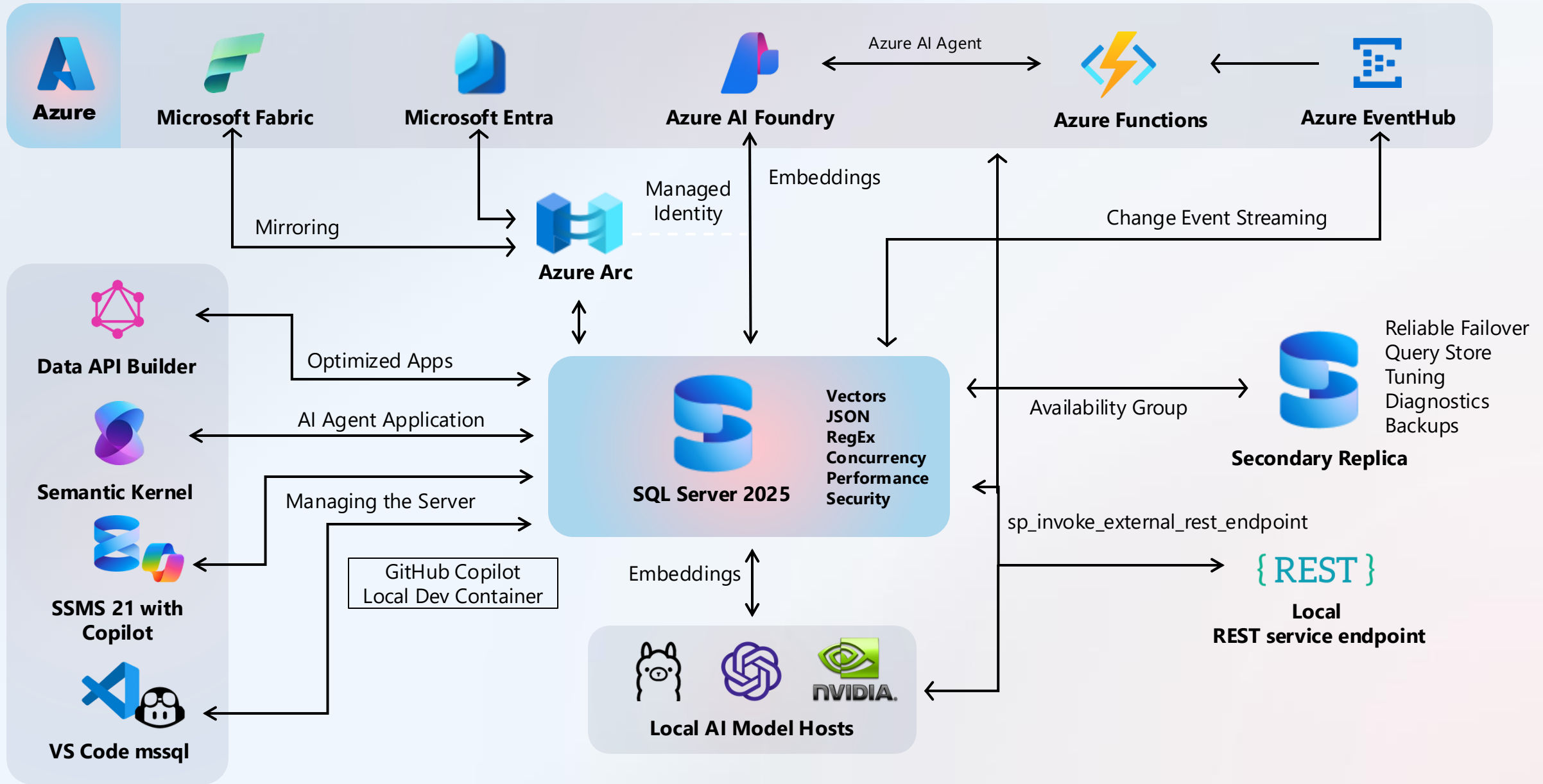
Tabelle Ledger per lo storico delle chat

# SQL Server 2025 Vector Search

**API\_TYPE**  
Azure OpenAI  
OpenAI  
Ollama



# SQL Server 2025: visione di insieme



# Ollama

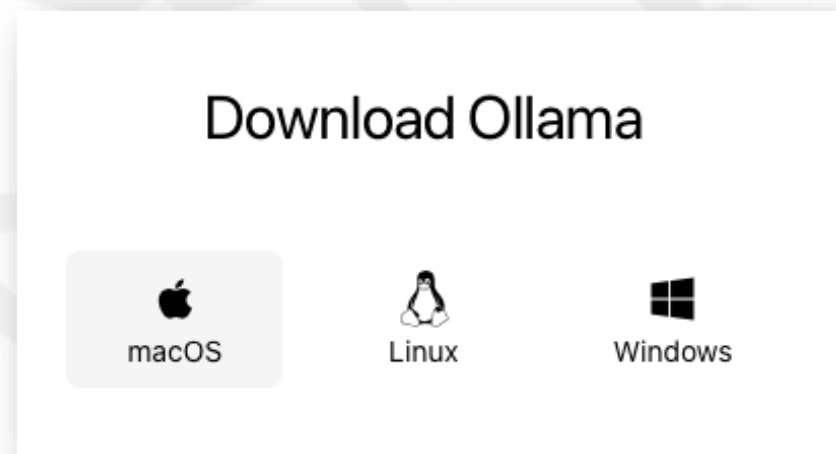


# Ollama

Piattaforma per eseguire *language models* localmente sul proprio computer

Può essere installato su ogni S.O.

Anche su Docker

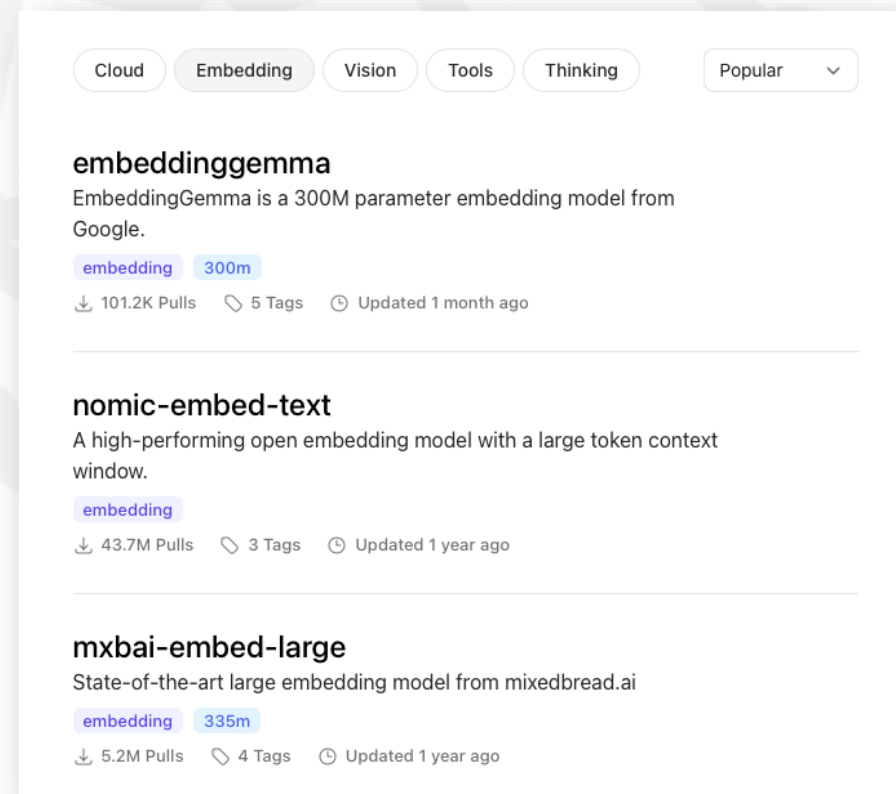




Scegliere il modello più adatto

Per task

Per utilizzo di memoria





# Scegliere il modello più adatto

Hugging Face mantiene una leaderboard dei modelli

- utile per identificare i migliori modelli

Rank (Boxed)	Model	Zero-shot	Memory Usage	Number of Parameters	Embedding Dimension	Max Tokens	Mean (Text)	Mean (Task)	Bitext	Classification	Clustering	Instruction R...
1	<a href="#">llama-embed-nemotron-8b</a>	99%	28629	7B	4096	32768	69.46	61.09	81.72	73.21	54.35	10.82
2	<a href="#">gemini-embedding-001</a>	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82	54.59	5.18
3	<a href="#">Qwen3-Embedding-8B</a>	99%	28866	7B	4096	32768	70.58	61.69	80.89	74.00	57.65	10.06
4	<a href="#">Qwen3-Embedding-4B</a>	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33	57.15	11.56
5	<a href="#">Qwen3-Embedding-0.6B</a>	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83	52.33	5.09
6	<a href="#">gte-Qwen2-7B-instruct</a>	⚠️ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55	52.77	4.94
7	<a href="#">Linq-Embed-Mistral</a>	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24	50.60	0.94
8	<a href="#">multilingual-e5-large-instruct</a>	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94	50.75	-0.40
9	<a href="#">embeddinggemma-300m</a>	99%	578	307M	768	2048	61.15	54.31	64.40	60.90	51.17	5.61
10	<a href="#">SFR-Embedding-Mistral</a>	96%	13563	7B	4096	32768	60.90	53.92	70.00	60.02	51.84	0.16
11	<a href="#">text-multilingual-embedding-002</a>	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64	47.84	4.08



# Ollama

## Client per interagire con Ollama

list

pull

show

serve

```
Usage:
  ollama [flags]
  ollama [command]

Available Commands:
  serve      Start ollama
  create     Create a model
  show       Show information for a model
  run        Run a model
  stop       Stop a running model
  pull       Pull a model from a registry
  push       Push a model to a registry
  signin     Sign in to ollama.com
  signout    Sign out from ollama.com
  list       List models
  ps         List running models
  cp         Copy a model
  rm         Remove a model
  help       Help about any command

Flags:
  -h, --help      help for ollama
  -v, --version    Show version information

Use "ollama [command] --help" for more information about a command.
```



# **Demo: SQL Server 2025 & Ollama vector search**



# Best practices



# Best practices

Usa modelli leggeri

Esegui inferenze in batch

Mantieni logging e tracciabilità delle chiamate AI

Monitoraggio performance e tempi di risposta

Utilizza cache dei risultati AI per ridurre costi e latenza

Redis



# Bilanciare Ollama con NGINX

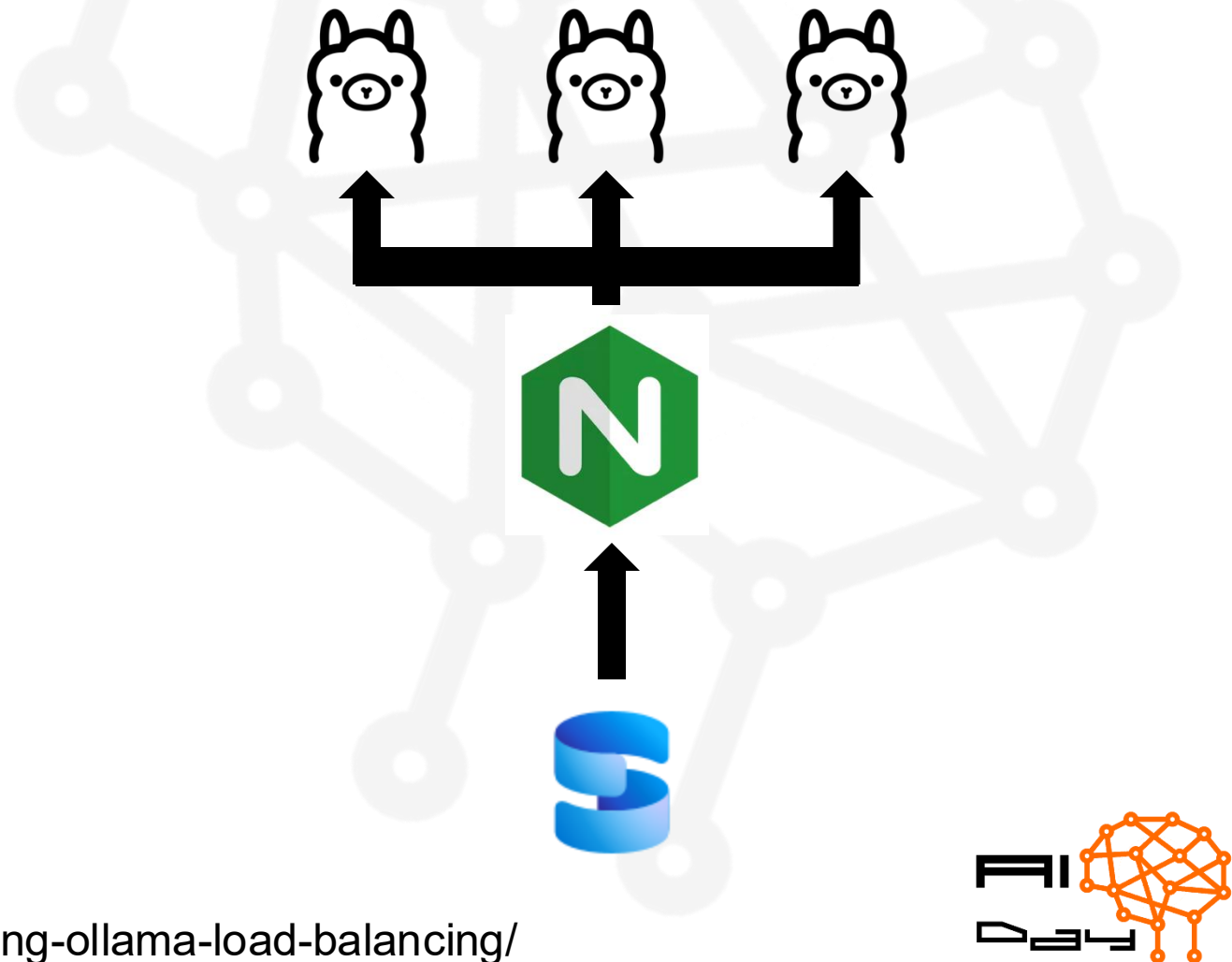
**Problema:** chiamare un servizio di embedding esterno a SQL Server via REST su HTTPs è limitato dalla banda del backend.

**Con dataset molto grandi i tempi di attesa possono essere molto lunghi**



# Bilanciare Ollama con NGINX

Soluzione: bilanciare le chiamate di generazione degli embedding utilizzando istanze multiple di Ollama bilanciate ad esempio da Nginx.





# Sicurezza e protezione dei dati

Come ogni software, anche Ollama può avere bug e falle di sicurezza:

- File Disclosure Vulnerabilities (CVE-2024-39722, CVE-2024-39719)
- Vulnerabilità dei modelli



# Sicurezza e protezione dei dati

Dare priorità nel disegnare l'architettura alla sicurezza:

- Principio di esposizione minima
- Verifica della disponibilità ed analisi degli aggiornamenti
- Accesso multilivello: combinare firewall, restrizione degli indirizzi IP e segmentazione di rete per minimizzare la possibilità di accesso non desiderato



# Demo: usare più istanze di Ollama bilanciate per ottimizzare i tempi

<https://github.com/nocentino/ollama-lb-sql>





QA

# Risorse

## What's new in SQL Server 2025 RC1

<https://learn.microsoft.com/en-us/sql/sql-server/what-s-new-in-sql-server-2025?view=sql-server-ver17>

## Scaling Ollama with load balancing

<https://www.nocentino.com/posts/2025-09-27-scaling-ollama-load-balancing/>



# Riferimenti

**Danilo Dominici**  
Senior consultant

@ ddominici@gmail.com

 [linkedin.com/in/danilodominici](https://www.linkedin.com/in/danilodominici)

 [github.com/ddominici/presentations](https://github.com/ddominici/presentations)



25 anni



2014-2020





Vote for this session!!