#GLOBALAZURE 2026

#Milano

Implementare Retrieval-Augmented Generation con Azure SQL e Microsoft Foundry

Andrea Saltarello

CEO @ Improove

#GLOBALAZURE
2026
#Milano

improove

TD SYNNEX

Grazie ai nostri sponsor 🙏

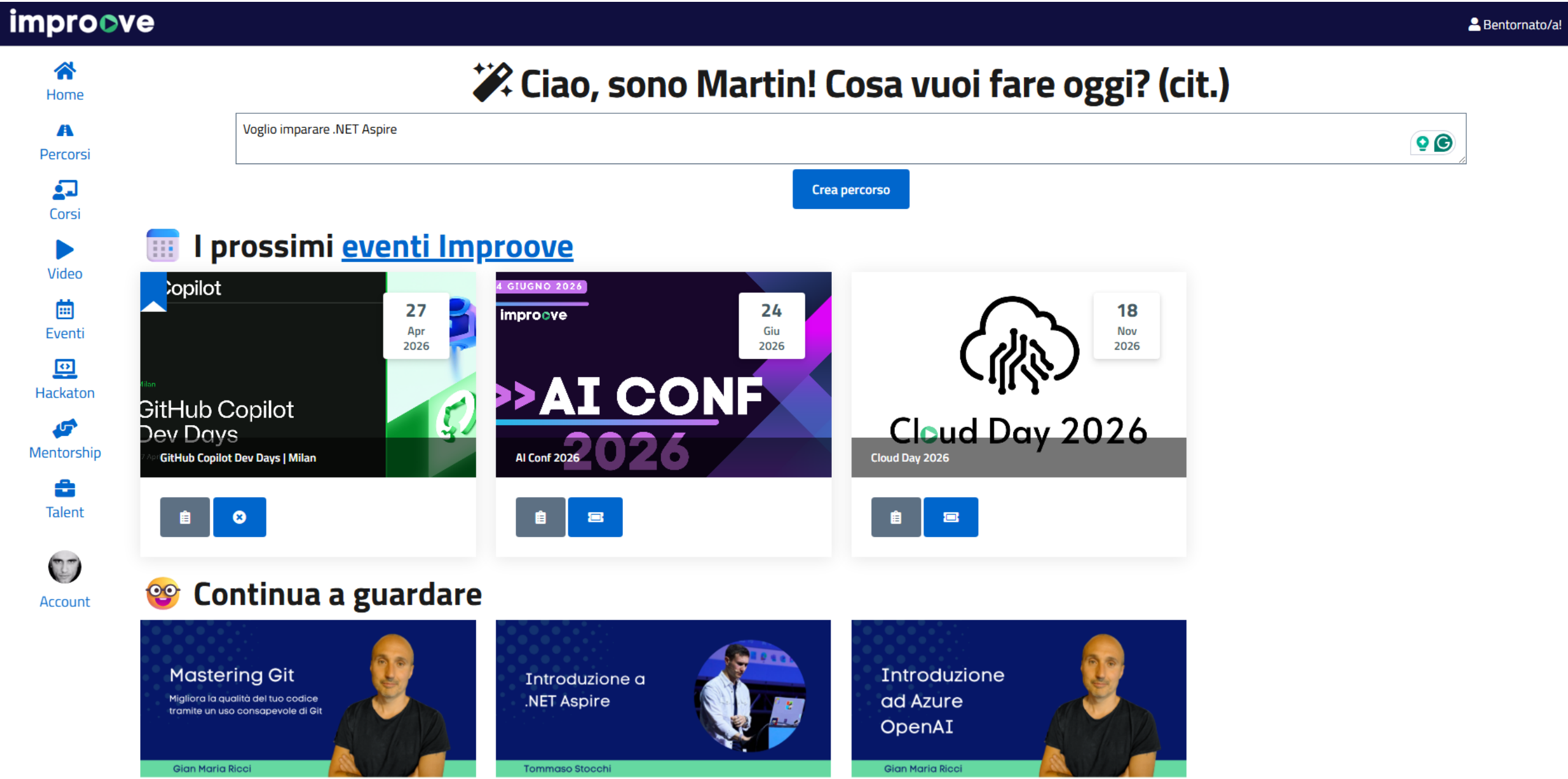
Contesto

Dalla ricerca vettoriale con Azure SQL ai modelli LLM e di embedding con Microsoft Foundry, Azure mette a disposizione tutti gli ingredienti necessari per implementare Retrieval-Augmented Generation: a fare la differenza, però, è la ricetta, a partire dall'arricchimento dei metadati.

In questo talk, vedremo l'implementazione di uno scenario reale.

There’s something about Martin

Il caso reale → http://www.improove.tech/app



GenAI basics

User input

The *text* written by the user

How much is a PS5?

Prompt

The *text* sent to the model in a given turn.

<|im_start|>system\n*You are an Xbox customer support agent whose primary goal is to help users with issues they are experiencing with their Xbox devices. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to Xbox.*\n**<|im_end|>**\n**<|im_start|>user**\n*How much is a PS5?*\n**<|im_end|>**\n**<|im_start|>assistant**

System Prompt

A fixed, application-controlled instruction block that establishes the model's identity, goals, behavioral rules, and response constraints. It is included at the beginning of the effective prompt and guides how the model should interpret all subsequent user inputs.

You are an Xbox customer support agent whose primary goal is to help users with issues they are experiencing with their Xbox devices. You are friendly and concise. You only provide factual answers to queries, and do not provide answers that are not related to Xbox.

Context window

The maximum number of tokens a language model can process at once, including both the input it receives and the output it generates. It defines how much text the model can “see” and use as context when producing its next response.

Tokens

Text chunk: «*The hyperquantized xenobiologist reexamined the microfractures.*»

Tokens: **The**, **hyper**, **quant**, **iz**, **ed**, **x**, **eno**, **bio**, **log**, **ist**, **re**, **exam**, **in**, **ed**, **the**, **micro**, **fract**, **ur**, **es**, **.**

Token IDs: 108, 2101, 103909, 286, 374, 302, 4024, 12971, 8292, 1117, 603, 19125, 374, 374, 279, 10641, 28759, 829, 315, 13

Llama 3: https://huggingface.co/meta-llama/Meta-Llama-3-Tokenizer

Llama 2: https://huggingface.co/meta-llama/Llama-2-7b-hf

RAG


1P
34200
ON I
50000
2P
3600
K.O
82
ChatGPT
Microsoft Copilot
SUPER

There’s more than meets the ~~eye~~ model

- If ChatGPT-the-model’s knowledge cut off date is August 31, 2025, why is it capable to provide info which weren’t available at that time?
- If ChatGPT-the-app and Copilot share the same underlying model, why do they answer differently?

KB-friendly prompts

You are given financial data for multiple companies. Use the provided fields - fiscal year, filing date, revenue, costs, EBITDA, and country - to compute the EBITDA-to-revenue ratio for each company.
Identify the top five companies operating in *Italy* based on this ratio.
Return the result as a ranked list with company name, ratio, and fiscal year.

Context data:

- Company A: FY 2023, filed 2024-03-12, revenue €120M, costs €95M, EBITDA €25M, country: Italy

- Company B: FY 2023, filed 2024-02-28, revenue €80M, costs €60M, EBITDA €20M, country: Italy

KB-friendly prompts (Reloaded)

You are given financial data for multiple companies. Use the provided fields - fiscal year, filing date, revenue, costs, EBITDA, and country - to compute the EBITDA-to-revenue ratio for each company. Identify the top five companies operating in *Italy* based on this ratio.
Return the result as a ranked list with company name, ratio, and fiscal year.

Context data:

- Company A: FY 2023, filed 2024-03-12, revenue €120M, costs €95M, EBITDA €25M, country: Italy
- Company B: FY 2023, filed 2024-02-28, revenue €80M, costs €60M, EBITDA €20M, country: Italy

Constraints:

- Use ONLY information present in the retrieved context
- If information is missing, state it explicitly

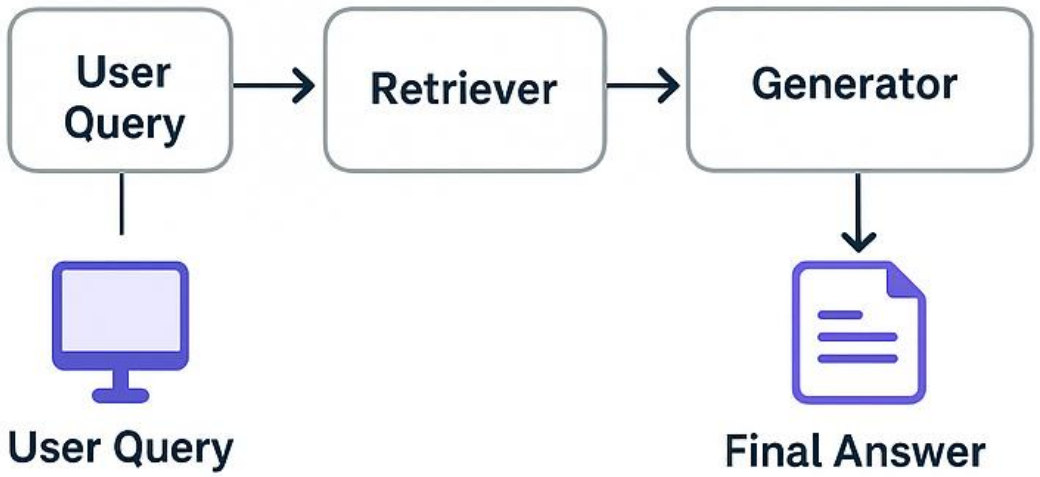
Embeddings

The numerical vector representations of text that capture semantic meaning

Text chunk: «*The hyperquantized xenobiologist reexamined the microfractures.*»

[-0.037388022989034 65, -0.014291535131633282, -0.00103846204001 4565, 0.021759940311312675, -0.024497266858816147, -0.000908682181 0342371, 0.058247581124305725, -0.005333060398697853, -0.043387819081544876, 0.062684245407581 33, -0.015484211966395378, -0.028439883142709732, -0.012881849892437458, 0.039105780422687 53, 0.050889678299427 03, ...]

How RAG works



1. **Retrieval Phase:** Fetches relevant information from a database or knowledge base.

2. **Augmentation Phase:** Injects retrieved data into the model's context.

3. **Generation Phase:** Produces an output based on both the prompt and retrieved information.

Implementing RAG

Ingestion – key points

- **Document normalization** - Convert heterogeneous sources (PDF, HTML, Office files, transcripts) into clean, machine-readable text.
- **Chunking** - Split long documents into semantically coherent segments that fit within the model's context window.
- **Embedding generation -** Convert each chunk into a vector representation capturing its semantic meaning.
- **Metadata enrichment** - Attach useful attributes (source, author, section, timestamp, tags) to improve retrieval precision.
- **Indexing** - Store embeddings and metadata in a vector database optimized for similarity search.

Augmentation – key points

- **Context assembly** - Merging the retrieved information into a coherent, model-ready context
- **Prompt grounding** - Anchoring the user request to verified, retrieved information so the model generates answers that stay aligned with factual context rather than relying solely on its internal priors.

Generation – key points

- **Style and tone control** - Adapting the response to the desired format, voice or level of detail
- **Hallucination mitigation** - Avoiding unsupported claims by prioritizing grounded, verifiable content
- **Output structuring** - Presenting the final answer in a clear, organized and user-appropriate format

S’el custa?

West Europe, no commitment

improove

RAG a la carte

- LLM (Llama @ Microsoft Foundry)
- Embeddings Model (Cohere @ Microsoft Foundry)
- Vector Database (Azure SQL)

S’el custa?

- Microsoft Foundry Prices:
  - LLM: input: 0,781 $ / 1M token
  - Cohere Embed: 0,1 $ per 1M token
  - OpenAI Whisper: 0,48 $ per hour
- Ingestion:
  - ~35M token LLM (input : output = ~12 : 1) ➡️ 24,85 USD
  - ~9M token Embeddings ➡️ 0,9 USD
- Inference
  - ~13K token LLM (input : output = ~20 : 1) ➡️ 0,0097 USD
- Vector Store:
  - 50 eDTU ➡️ ~110.27 USD/month
- Speech To Text:
  - ~~480$~~ Locally done with an NVIDIA 5070 Ti

CYBERPUNK 2077

ULTIMATE EDITION

#GLOBALAZURE
2026
#Milano

Slide e video:
https://www.globalazuremilano.it

Andrea Saltarello

https://www.linkedin.com/in/andysal/
https://github.com/andysal