

Dal feedback umano ai dati: RAG evaluation senza sorprese

26 02 2026

Agenda

 **1.**

Introduzione al RAG

 **2.**

Contesto e sfide

 **3.**

Aletheia: evaluation framework per RAG

 **4.**

Demo

 **5.**

Reasoning: costi e tempi

Introduzione al RAG (1/2)



Limitazioni degli LLMs

- Training fermo a un dato **momento storico**
- Addestramento da 0 **costoso** dal punto di vista **computazionale**



Il problema

- Come può il modello basarsi su **informazioni aggiornate** nel tempo?
- Come può accedere ad **informazioni interne** all'azienda?

Introduzione al RAG (1/2)



Limitazioni degli LLMs

- Training fermo a un dato **momento storico**
- Addestramento da 0 **costoso** dal punto di vista **computazionale**

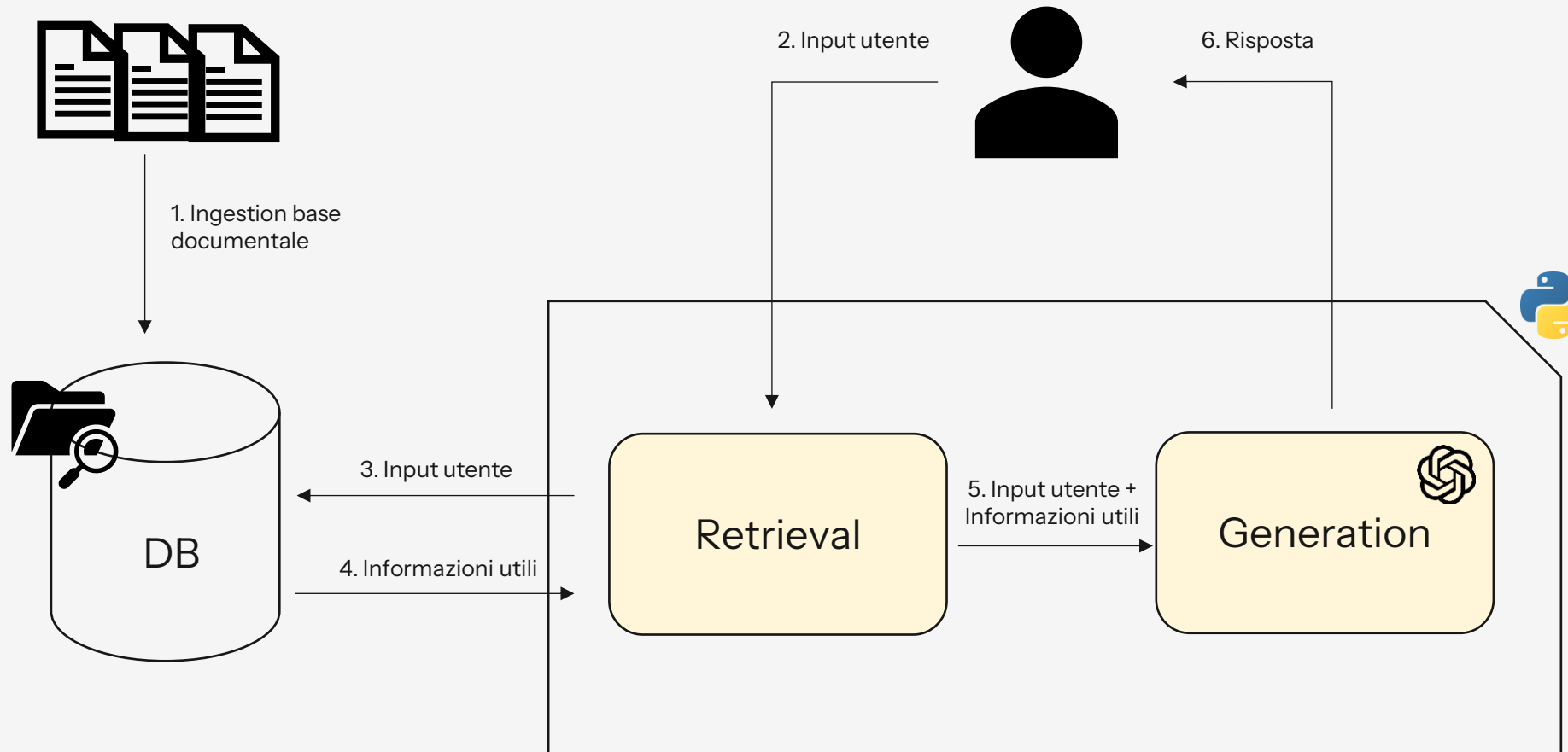


Il problema

- Come può il modello basarsi su **informazioni aggiornate** nel tempo?
- Come può accedere ad **informazioni interne** all'azienda?

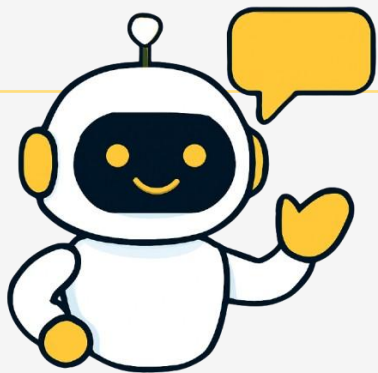
RAG

Introduzione al RAG (2/2)



Contesto e sfide (1/3)

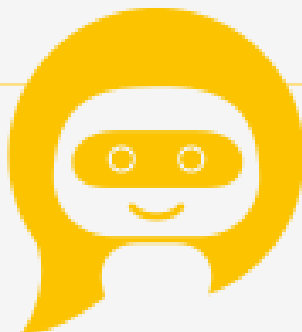
Guide Utili



«**Accesso** intelligente e **conversazionale** alle Guide Utili di MyFastweb per migliorare l'esperienza di **assistenza del cliente**»

Tra 350 e 400
sessioni al giorno

CHIARA



«Un assistente virtuale basato su GenAI che **semplifica** l'accesso a **policy e processi aziendali**, direttamente da intranet e **Teams**»

Più di 500
conversazioni al giorno

OneDesk



«Un Help Desk più efficiente che sfrutta la GenAI e l'**Agentic Orchestration** per risolvere i **problemi dei dipendenti**»

1st Agentic tool
introdotto in Fastweb

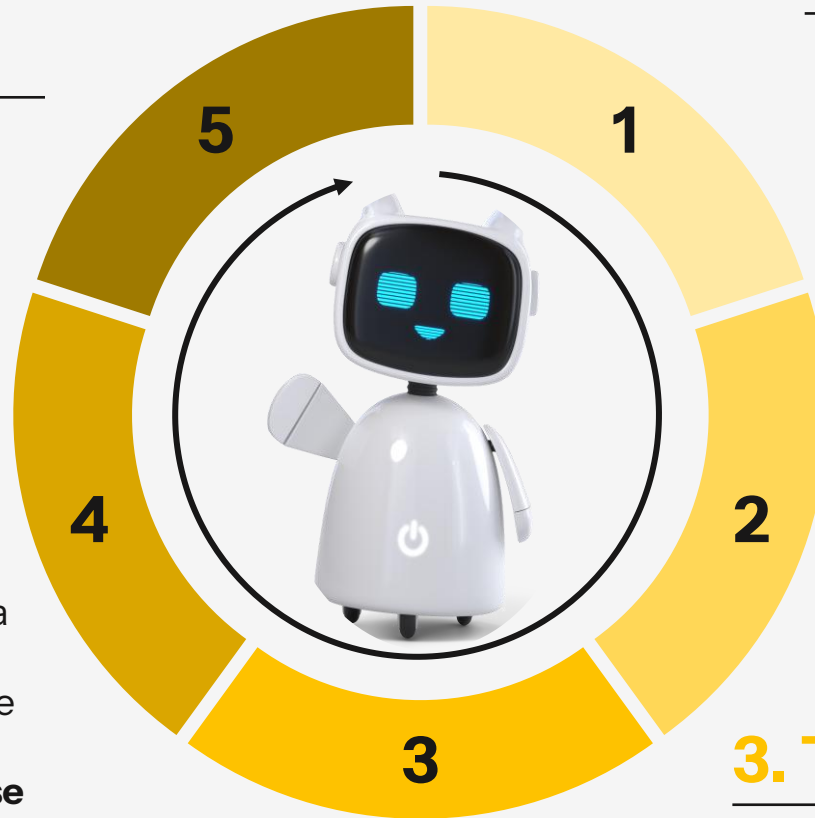
Contesto e sfide (2/3)

5. Monitoraggio

Raccogliere i feedback degli utenti è essenziale per **definire al meglio i prossimi sviluppi**

4. Deploy

Dopo aver **ottimizzato** la soluzione sulla base dei **feedback** dei pionieri, ogni chatbot è stato rilasciato in produzione e **messo a disposizione dell'intera popolazione aziendale o customer base**



1. Design

Individuazione dei need e design di una soluzione che ottimizzi l'esperienza dell'utente

2. Sviluppo

Realizzazione e configurazione del chatbot **secondo i requisiti definiti**, con integrazione nei sistemi aziendali e preparazione al testing

3. Test

Ogni chatbot sviluppato è stato testato da un gruppo di dipendenti (**pionieri**) i cui feedback hanno contribuito a **migliorarne le funzionalità**

Contesto e sfide (3/3)



La valutazione umana dei sistemi RAG è **time-consuming** e **soggettiva**



Standardizzazione, automazione e oggettività nella valutazione delle performance



Mancanza di un framework per la valutazione quantitativa dell'efficacia dei sistemi RAG



Necessità di **monitoraggio continuo** della qualità delle risposte del chatbot **in produzione**

Contesto e sfide (3/3)



La valutazione umana dei sistemi RAG è **time-consuming** e **soggettiva**



Standardizzazione, automazione e oggettività nella valutazione delle performance



Mancanza di un framework per la valutazione quantitativa dell'efficacia dei sistemi RAG



Necessità di **monitoraggio continuo** della qualità delle risposte del chatbot **in produzione**

Aletheia!

Aletheia: evaluation framework per RAG



Aletheia è un **pacchetto Python*** che consente di effettuare una valutazione oggettiva, quantitativa e automatizzata dei **sistemi RAG**



Il tool misura la qualità dei **chunk recuperati** (fase di *retrieval*) e delle **risposte generate** dai LLM (fase di *generation*) sulla base di **metriche oggettive** tramite l'uso di un **LLM-as-a-judge** di Azure OpenAI, AWS Bedrock o Google



È pensato per lo sviluppo e il miglioramento degli use-cases RAG-based, fornendo metriche utili per il **tuning**, la **validazione** e il **monitoraggio continuo** delle soluzioni nel tempo

* Attualmente è un wrapper del pacchetto Python RAGAS

Metriche di valutazione (1/5)

RETRIEVAL

-  Selezionare **documenti pertinenti**
-  Minimizzare **informazioni irrilevanti**



Recuperare il **contesto giusto**



Metriche:

- Context Precision (reference)
- Context Precision (no reference)
- Context Recall

Metriche di valutazione (1/5)

RETRIEVAL

- 🎯 Selezionare **documenti pertinenti**
- 🎯 Minimizzare **informazioni irrilevanti**



Recuperare il **contesto giusto**



Metriche:

- Context Precision (reference)
- Context Precision (no reference)
- Context Recall

GENERATION

- 🎯 Risposta **coerente e fedele** al contesto
- 🎯 Evitare **allucinazioni**



Generare la **risposta corretta**



Metriche:

- Faithfulness
- Factual Correctness
- Rubrics

Metriche di valutazione (2/5) - RETRIEVAL

Context Precision With Reference



A cosa serve?

*Per valutare quanti **chunks recuperati** sono utili*

Dati necessari per il calcolo

- ✓ User input
- ✓ Reference
- ✓ Retrieved Chunks

Metriche di valutazione (2/5) - RETRIEVAL

Context Precision With Reference



A cosa serve?

Per valutare quanti **chunks recuperati** sono utili

Dati necessari per il calcolo

- ✓ User input
- ✓ Reference
- ✓ Retrieved Chunks

Context Precision Without Reference



A cosa serve?

Per valutare quanti **chunks recuperati** sono utili

Dati necessari per il calcolo

- ✓ User input
- ✓ Response
- ✓ Retrieved Chunks

Metriche di valutazione (2/5) - RETRIEVAL

Context Precision With Reference



A cosa serve?

Per valutare quanti **chunks recuperati** sono utili

Dati necessari per il calcolo

- ✓ User input
- ✓ Reference
- ✓ Retrieved Chunks

Context Precision Without Reference



A cosa serve?

Per valutare quanti **chunks recuperati** sono utili

Dati necessari per il calcolo

- ✓ User input
- ✓ Response
- ✓ Retrieved Chunks

Context Recall



A cosa serve?

Per valutare quanti tra i **chunks utili in Knowledge Base** sono stati recuperati

Dati necessari per il calcolo

- ✓ User input
- ✓ Reference
- ✓ Retrieved Chunks

Metriche di valutazione (3/5) - RETRIEVAL



Where is the Eiffel Tower located?

The Eiffel Tower is located in Egypt



Reference

"The Eiffel Tower is located in Paris"

**Retrieved
chunks**

[*"The Eiffel Tower is one of the most famous landmarks in Paris."*,
"Rome is the capital of Italy."]

Metriche di valutazione (3/5) - RETRIEVAL



Where is the Eiffel Tower located?

The Eiffel Tower is located in Egypt



Reference

“The Eiffel Tower is located in Paris”

**Retrieved
chunks**

[“The Eiffel Tower is one of the most famous landmarks in Paris.”,
“Rome is the capital of Italy.”]

METRICHE

Context Precision
With Reference

0.5

Context Precision
Without Reference

0

Context Recall

1

Metriche di valutazione (4/5) - GENERATION

Faithfulness



A cosa serve?

Per valutare il grado di **allucinazioni** prodotte

Dati necessari per il calcolo

- ✓ User input
- ✓ Response
- ✓ Retrieved Chunks

Metriche di valutazione (4/5) - GENERATION

Faithfulness



A cosa serve?

Per valutare il grado di **allucinazioni** prodotte

Dati necessari per il calcolo

- ✓ User input
- ✓ Response
- ✓ Retrieved Chunks

Factual Correctness



A cosa serve?

Per valutare quanto la **risposta** è **fattuale** rispetto alla **ground truth**

Dati necessari per il calcolo

- ✓ Reference
- ✓ Response

Metriche di valutazione (4/5) - GENERATION

Faithfulness



A cosa serve?

Per valutare il grado di **allucinazioni** prodotte

Dati necessari per il calcolo

- ✓ User input
- ✓ Response
- ✓ Retrieved Chunks

Factual Correctness



A cosa serve?

Per valutare quanto la **risposta** è **fattuale** rispetto alla **ground truth**

Dati necessari per il calcolo

- ✓ Reference
- ✓ Response

Rubrics



A cosa serve?

Per valutare se la risposta soddisfa una **descrizione di requisiti**

Dati necessari per il calcolo

- ✓ Reference
- ✓ Response

Metriche di valutazione (5/5) - GENERATION



Where is the Eiffel Tower located?

The Eiffel Tower is located in Egypt



Reference

"The Eiffel Tower is located in Paris"

**Retrieved
chunks**

[*"The Eiffel Tower is one of the most famous landmarks in Paris."*,
"Rome is the capital of Italy."]

Metriche di valutazione (5/5) - GENERATION



Where is the Eiffel Tower located?

The Eiffel Tower is located in Egypt



Reference

“The Eiffel Tower is located in Paris”

Retrieved
chunks

[“The Eiffel Tower is one of the most famous landmarks in Paris.”,
“Rome is the capital of Italy.”]

METRICHE

Faithfulness

0

Factual
Correctness

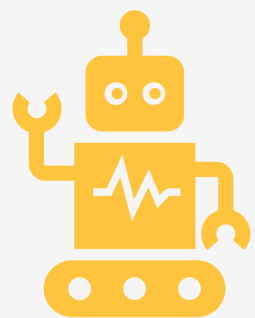
0

Rubrics

1

Rubrics description:

- **score 1:** Completely incorrect and does not address the reference.
- **score 2:** Partially correct but contains major errors or omissions.
- **score 3:** Mostly correct but lacks clarity or important details.
- **score 4:** Accurate and clear with only minor issues.
- **score 5:** Fully accurate, clear, and complete.



Demo!

Reasoning: costi e tempi



Reasoning: costi e tempi



- ✓ **Progetto:** *Guide Utili*
- ✓ **Numero Q&A:** 80
- ✓ **Modello:** o3
- ✓ **Costo totale:** \$15.17
- ✓ **Tempo totale:** 4 h 5 min

Conclusioni

PRIMA

Valutazione soggettiva

Procedura lunga e complessa

Feedback umano discontinuo



DOPO

Valutazione oggettiva

Processo strutturato ed efficiente

Monitoraggio quantitativo e continuo



FASTWEB +  **vodafone**