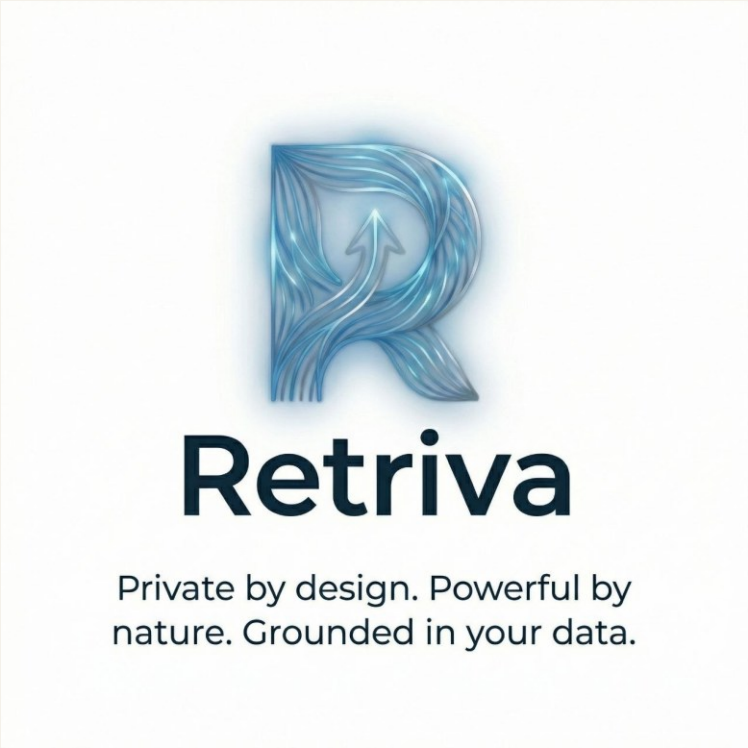
CLOUD CONFERENCE ITALIA

10ª EDIZIONE

Agente conversazionale per uso aziendale/professionale basato su RAG e infrastruttura *confidential computing*

Retriva

CHI SONO



SPEAKER

Andrea Marson

Strategic Technical Advisor, DAVE Srl

DAVE Embedded Systems

Bio

- 25+ anni di esperienza nel settore dei sistemi embedded, maturati in DAVE Embedded Systems, in cui ho ricoperto vari ruoli:
  - Ingegnere dei sistemi embedded
  - CTO
  - Membro del consiglio di amministrazione
  - Advisor del il reparto commerciale per attività di technical marketing
  - Questo mi ha permesso di vedere il business da molteplici prospettive diverse
- A distanza di tempo dalla laurea in Ingegneria Elettronica, ho iniziato a frequentare dei master in Intelligenza Artificiale
- Nel « tempo libero», investitore in start-up e co-founder

Introduzione

Origine del progetto

- Trovare un progetto per il master in AI iniziato nel 2024 che consentisse di mettere in pratica (parte di) quanto appreso in uno scenario reale o quantomeno realistico
- Casus belli: necessità reali sperimentate in azienda, prima fra tutte il bisogno di reperire efficacemente le informazioni in un sito interno realizzato con *Mediawiki*, contenente documentazione **riservata** redatta in decenni di lavoro

È possibile individuare un progetto che coniughi questi due punti?



Retriva

PS Perché non usare qualcosa di già disponibile nel mercato? All'epoca, non avevo trovato nulla in grado di soddisfare i requisiti.

Cos’è

- Tecnicamente, è un agente conversazionale con funzionalità *retrieval-augmented generation* (RAG)
- Da un punto di vista funzionale, si presenta quindi all’utente fondamentalmente come un chatbot in stile ChatGPT
- Pensato per essere utilizzato in ambiente aziendale/professionale, quindi garantendo
  - **Accuratezza dell’output** (meglio una non risposta di una risposta sbagliata)
  - **Protezione dei dati aziendali** che si assume siano per default di **natura**
- Grazie alla sua espandibilità, può integrare anche funzionalità di assistente digitale per svariate funzioni aziendali

Knowledge base (base di conoscenza)

- La *Knowledge Base* (KB) è un **concetto fondamentale** nei sistemi RAG
- Rappresenta la conoscenza sulla base della quale il chatbot produce il suo output
- Più prosaicamente, è l'«archivio» dei dati da cui il chatbot, attraverso un opportuno processo di *retrieval*, estrae le informazioni «giuste» per costruire le proprie risposte, generare documenti, prendere decisioni, ecc.
- Nelle aziende di solito questa conoscenza è sparpagliata in tanti documenti di natura diversa: file PDF, file Word, fogli elettronici in formato Excel, registrazioni audio/video, immagini, ecc. Retriva è in grado di **«digerire» questi documenti per estrarne le informazioni rilevanti e riversarle nella propria base di conoscenza**.

Obiettivi

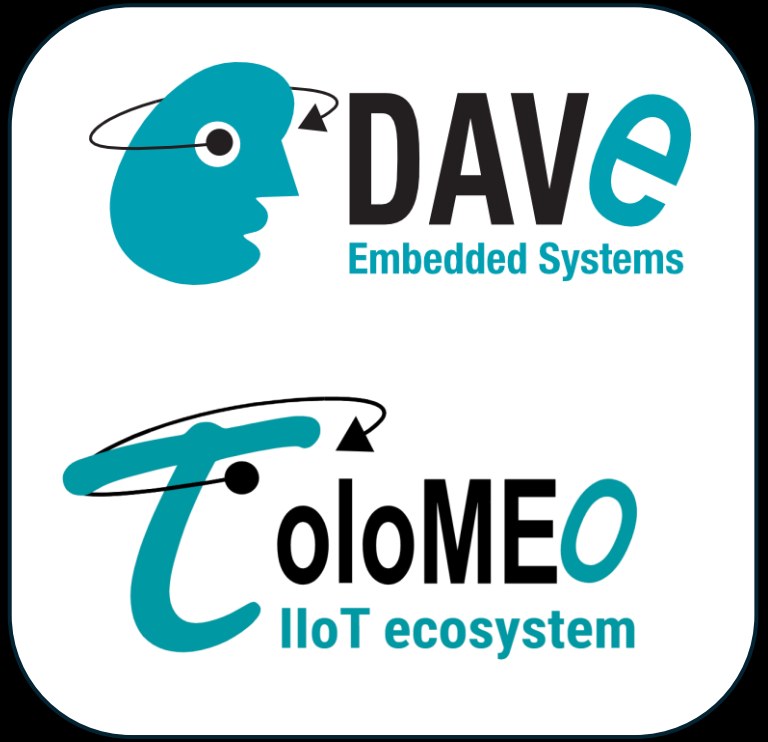
A partire da quanto detto, mi sono quindi posto i seguenti obiettivi:

- Sviluppare da zero un agente conversazionale con le caratteristiche riassunte precedentemente, utilizzando esclusivamente un approccio *Spec-Driven Development* (SDD)
- Ottenere un software sufficientemente strutturato da poter essere usato in un contesto aziendale reale e — perché no? — da poter evolvere un domani in un prodotto vero e proprio
- Esplorare soluzioni alternative per affrontare il problema della protezione dei dati
- Sperimentare diversi tool e LLM per il coding automatico

Alcuni casi d’uso

Contesto: DAVE Embedded Systems

- Progettazione e produzione di sistemi embedded per applicazioni industriali (automazione, biomedicale, trasporti, telecomunicazioni, aerospazio, difesa, ecc.)
- Ecosistema AIoT ToIoMEO
  - Gestione flotta dispositivi edge in ambito industriale e OT
  - Aggiornamenti sicuri OTA
  - Gestione SBOM (compliance con normativa Cyber Resilience Act)
  - Troubleshooting remoto
  - AI at the edge
  - …

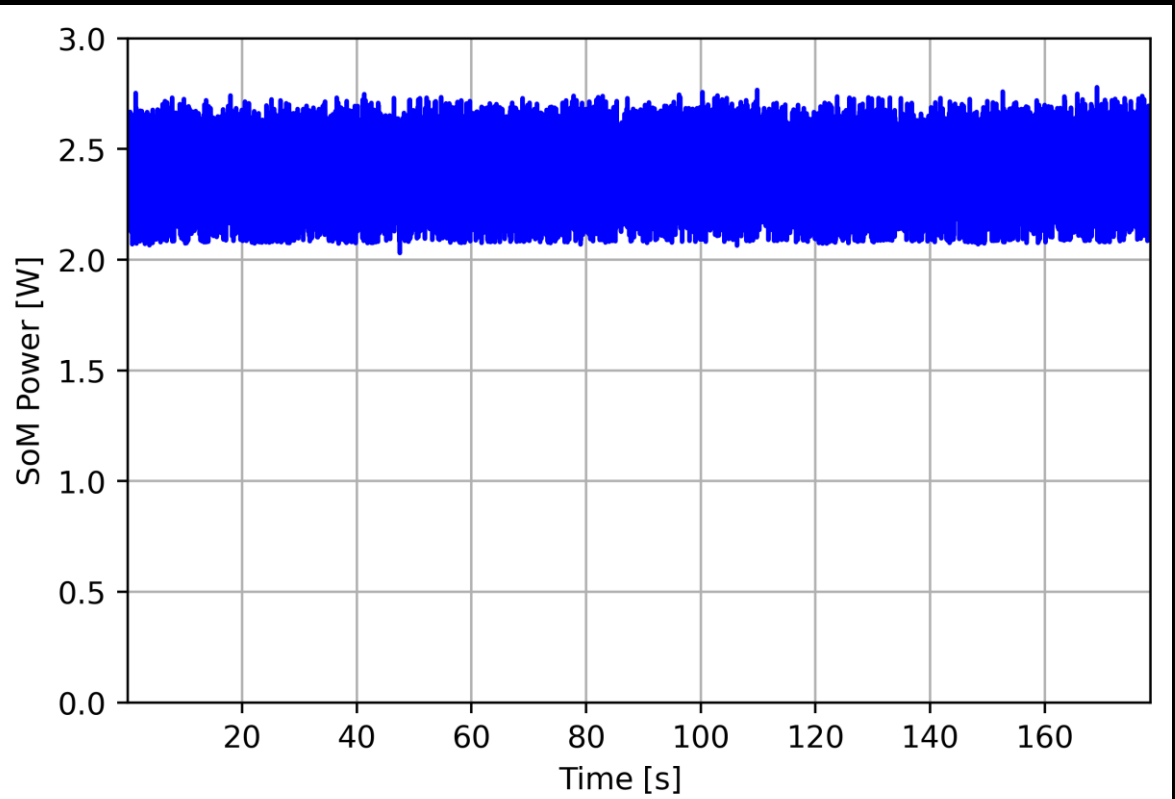


Ricerca avanzata tra sorgenti eterogenee

What is the maximum power consumption of AURA?

- AURA è una scheda elettronica i cui consumi sono stati caratterizzati sperimentalmente in svariate condizioni di utilizzo diverse
- Rispondere a questa domanda non è banale, tenendo conto del fatto che la risposta non è presente in forma esplicita nella *knowledge base* e va ricavata estraendo le misure (rappresentate in forma testuale, tabellare, o grafica) da diverse sorgenti → vedi slide seguente

ID	Mode group	Mode	Cortex A55	Cortex M33	SDRAM [MT/s]	VDD_SOC [V]	Power [W]		
							min	avg	max
1	Run mode	Over drive	• 1.7 GHz • No specific user space applications were run	• reset	3733	0.90	1.601	**1.696**	2.211
2		Nominal	• 1.4 GHz • No specific user space applications were run	• reset	2800	0.85	1.543	**1.584**	1.838
			• 933 MHz						



retriva + ··· AM

What is the maximum power consumption of AURA SOM?

retriva Today at 4:29 PM

Based on the provided context, there is no single explicit "maximum power consumption" specification (e.g., a rated P_max) listed for the AURA SOM in the text snippets. However, measured power consumption values from various operational modes and stress tests are documented in the image descriptions within the **AURA SOM/AURA Hardware/pdf** source.

According to these measurements:

- **Stress Conditions:** Under stress conditions (specifically the `AURA_power_a55_stress_OD` test), the maximum observed power consumption is approximately **2.75 W to 2.8 W** [10]. Other stress tests (`AURA_power_a55_stress_ND` and `AURA_power_a55_stress_LD`) show maximums around **2.15 W** and **2.0 W** respectively [10].
- **Idle Conditions:** In idle states, peak power can reach approximately **2.2 W** (e.g., `AURA_power_a55_idle_OD`), with average idle power typically around **1.5 W to 1.7 W** [10].
- **Low Power Modes:** In low-power idle modes (`AURA_power_lp_idle`), consumption is around **0.6 W**. In suspend mode (`AURA_power_lp_suspend`), it drops to approximately **0.16 W**, and in BBSM mode (`AURA_power_lp_bbsm`), it is approximately **1.10 mW** [10].

The AURA SOM is based on the i.MX93 SoC, which supports innovative power management [1]. While a specific theoretical maximum rating is not provided in the text, the highest measured value in the provided stress test data is approximately **2.8 W**.

Note: A separate source titled "Thermal tests and measurements (SBCX)" mentions a maximum of 3.8 W, but the context does not explicitly confirm that the SBCX board uses the AURA SOM, so this value is not attributed to AURA in the provided text.



Studio professionale

La gestione della consulenza delle pratiche in uno studio professionale è complessa a causa delle normative in continuo aggiornamento. Retriva può offrire un accesso «smart» a normative sempre aggiornate, nonché allo storico dello studio.



Assistenza on-site

Intervenire in loco per operazioni di manutenzione/assistenza su macchine di una certa complessità non è facile, soprattutto quando l'operatore ha poca esperienza. Retriva può guidare il tecnico in queste operazioni, fornendo delle indicazioni mirate e facili da seguire, evitando di dover consultare tonnellate di documentazione per trovare le informazioni richieste.

CRM «intelligente»

Il valore «nascosto» nella KB aziendale relativamente ai rapporti con i clienti **è in potenza elevatissimo**. Un sistema come **Retriva può scovarlo autonomamente**, suggerendo di conseguenza **proattivamente** al personale commerciale le azioni più opportune da fare per sfruttarlo appieno.

Alcuni esempi di possibili funzionalità:

- *Dormant lead resurrection*
- *Next-best-action recommendation*
- *Promise and commitment tracking*
- *Account health monitoring and churn and dissatisfaction risk detection*
- *Cross-sell and upsell recommendations*
- *Support-to-sales opportunity detection*
- *Leads qualification* (maggiori dettagli in seguito)
- …

Un altro aspetto interessante che vale la pena sottolineare è che **funzioni come queste non necessariamente hanno bisogno del sottosistema RAG**, ma certamente beneficiano moltissimo dell'LLM usato per la chat in termini di semplicità d'uso per l'utente.

Sistemi RAG: alcuni richiami

Sistemi RAG: alcuni richiami (1)

L'operatività dei sistemi RAG si articola sostanzialmente in due processi:

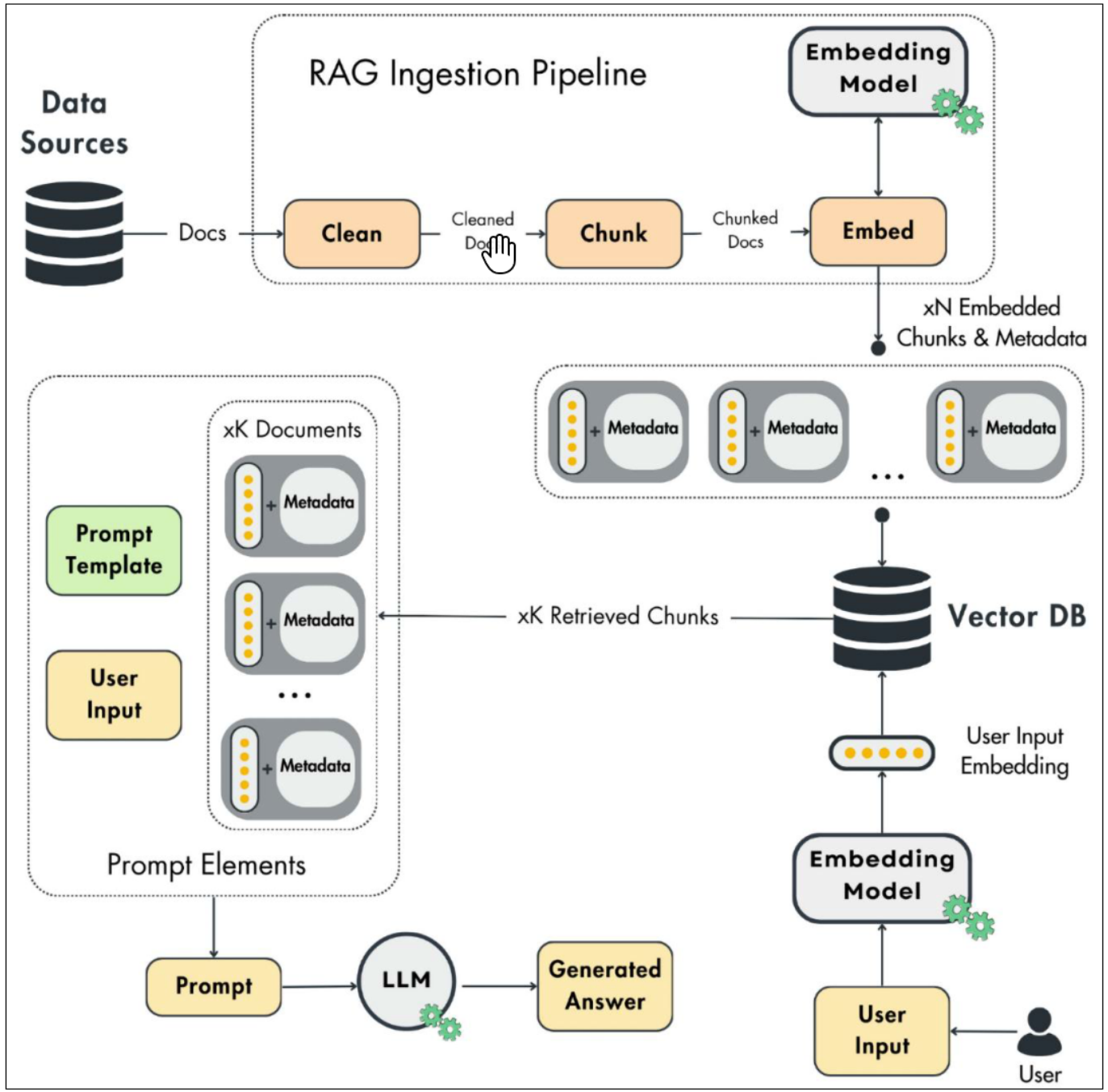
- L'inserimento nella *Knowledge Base* dei dati di interesse (si parla di *ingestion* dei documenti)
- L'utilizzo della conoscenza inclusa nella KB (da parte di utenti o di agenti) attraverso un chatbot o altri strumenti che sfruttano un LLM che opera dietro le quinte
  - Questo consente di estendere la conoscenza intrinseca dell'LLM (*parametrized knowledge*) **senza alcuna (generalmente costosa) operazione di training**.
- La bontà di un RAG è determinata principalmente dalla qualità di
  - Chunking/embedding
  - Retrieval

Sistemi RAG: alcuni richiami (2)

Perché si parla di generazione *retrieval-augmented*?

Un chatbot «puro» privo di RAG è in grado di rispondere esclusivamente sulla base della conoscenza insita nell'LLM che utilizza. Nel caso in cui gli chiedessimo, ad esempio, delle informazioni su un evento accaduto successivamente alla data di *cut-off* del relativo training, non avrebbe materialmente gli elementi per comporre una risposta coerente. Se dispone di funzionalità RAG, invece, può reperire (*retrieval*) questi dati mancanti (tipicamente accedendo ad Internet, ma anche ad una sorgente locale) ed operare quindi una generazione arricchita (*augmented*) della risposta (1), combinando la conoscenza intrinseca dell'LLM con quella aggiuntiva scaturita dal processo di *retrieval*.

(1) Ciò che viene tecnicamente arricchito è il prompt dato in pasto all'LLM sottostante, fornendo in questo modo un contesto con informazioni addizionali.

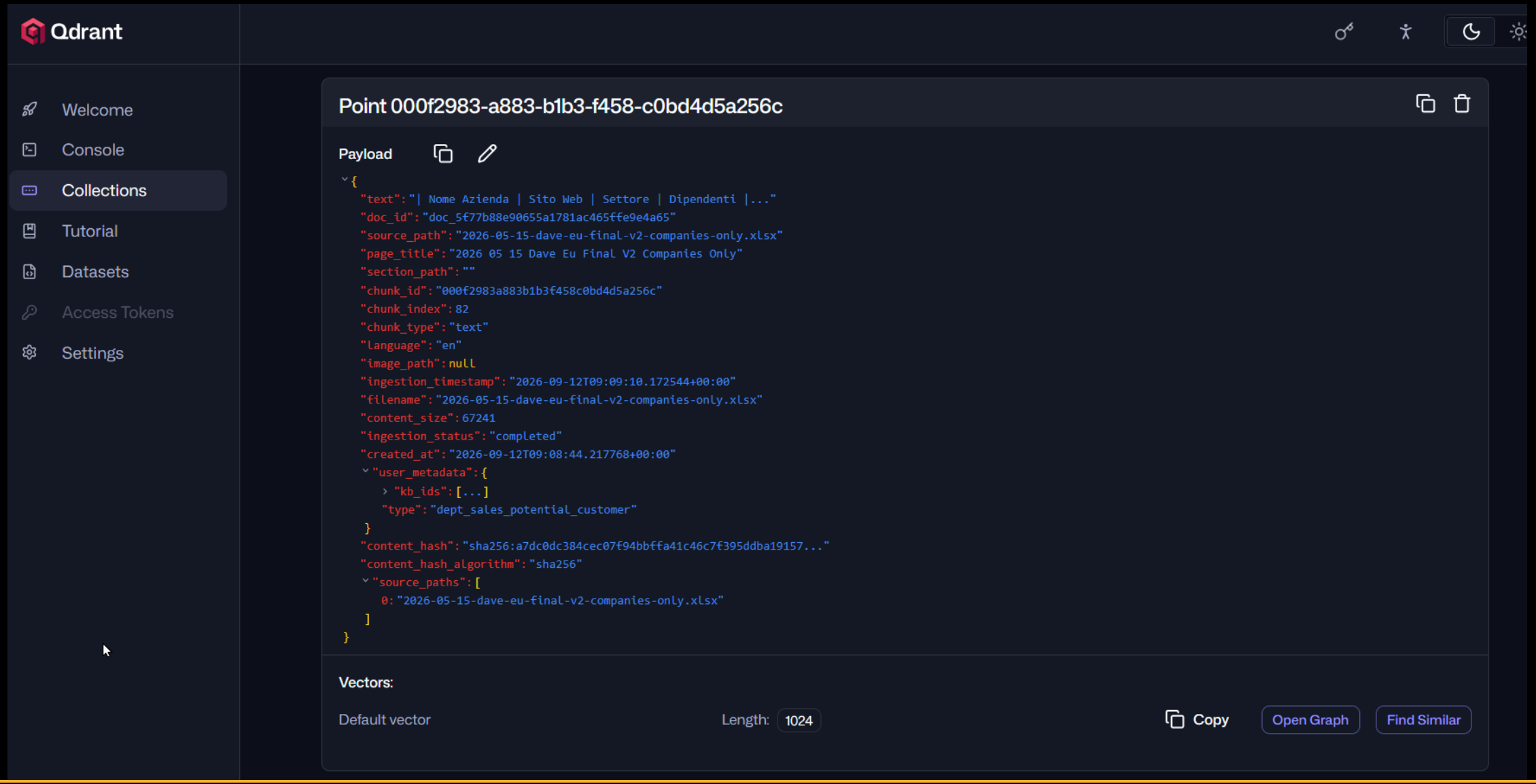


Sistemi RAG: alcuni richiami (3)

Architettura tipica

Source:
https://github.com/AkshaysNimbalkar/LLM-Engineers-Handbook/blob/master/Handbbok_Color_Images.pdf

Sistemi RAG: alcuni richiami (4)



Sistemi RAG: alcuni richiami (5)

cust_0007



Code | Data Panel

Point 3151b2e1-ad1e-fa33-30e9-4eb4c495e83c

Payload


```
{
  "text": "| Nome Azienda | Sito Web | Settore | Dipendenti |..."
  "doc_id": "doc_5f77b88e90655a1781ac465ffe9e4a65"
  "source_path": "2026-05-15-dave-eu-final-v2-companies-only.xlsx"
  "page_title": "2026 05 15 Dave Eu Final V2 Companies Only"
  "section_path": ""
  "chunk_id": "3151b2e1ad1efa3330e94eb4c495e83c"
  "chunk_index": 99
  "chunk_type": "text"
  "language": "en"
  "image_path": null
  "ingestion_timestamp": "2026-09-12T09:09:10.172544+00:00"
  "filename": "2026-05-15-dave-eu-final-v2-companies-only.xlsx"
  "content_size": 67241
  "ingestion_status": "completed"
  "created_at": "2026-09-12T09:08:44.217768+00:00"
  "user_metadata": {
    "kb_ids": [...]
    "type": "dept_sales_potential_customer"
  }
  "content_hash": "sha256:a7dc0dc384cec07f94bbffa41c46c7f395ddba19157..."
  "content_hash_algorithm": "sha256"
  "source_paths": [
    0: "2026-05-15-dave-eu-final-v2-companies-only.xlsx"
  ]
}
```

Design

Requisiti fondamentali (1)

Premessa importante: Retriva nasce principalmente per trattare **documentazione tecnico/scientifica** utilizzata in ambito ingegneristico => i documenti da maneggiare sono relativamente complessi (tabelle, immagini ricche di informazioni, ecc.).

- Filosofia generale di progetto: essere ***data-centric*** e ***model-agnostic***
- Aspetti funzionali
  - Accuratezza dell'output generato da cui un approccio molto conservativo
    - L'output deve essere costruito a partire **esclusivamente** dalla KB
    - Se non si dispone di sufficienti dati per fornire un output coerente, il sistema deve esplicitarlo piuttosto che inventarsene uno plausibile (riduzione della probabilità di allucinazione) ovvero è meglio non avere risposte che averne una sbagliata
  - Grounding: fornire evidenza esplicita dei documenti a partire dai quali è stata costruita la risposta
  - Quasi determinismo (= a parità di domanda, ottengo la stessa risposta), ovvero ridurre al minimo la «creatività» dell'LLM sottostante
  - Supporto multilingua (inglese e italiano; possibili ulteriori lingue in futuro)
  - UX: ChatGPT-like

Requisiti fondamentali (2)

- Aspetti implementativi
  - Forte modularità per:
    - Favorire la manutenibilità
    - Avere flessibilità nell'estendere il sistema con nuove funzionalità
    - Facilitare il deployment in scenari diversi (on-prem, cloud, ibrido)
- Protezione dei dati
  - Prevedere fin da subito la possibilità di integrare feature specifiche per soddisfare questo requisito

Requisiti: protezione dei dati (1)

Alcuni casi di data leak (involontari?) causati dai maggiori provider AI

- Samsung Electronics (aprile 2023): Alcuni dipendenti hanno incollato per errore codice sorgente proprietario e verbali di riunioni riservate su ChatGPT per correggere bug o riassumere i testi. Essendo i dati finiti sui server di OpenAI per l'addestramento, l'azienda ha dovuto vietare l'uso di IA generativa sui dispositivi aziendali per il rischio che tali segreti venissero riproposti ad altri utenti.
- Bug della cronologia di ChatGPT (marzo 2023): Un problema tecnico ha permesso ad alcuni utenti di visualizzare i titoli e i primi messaggi delle conversazioni di altre persone. Se un utente aveva inserito dati sensibili nel primo prompt, questi sono rimasti visibili a estranei per diverse ore.
- Conversazioni indicizzate su Google (agosto 2025): È stato scoperto che migliaia di conversazioni private di ChatGPT erano state indicizzate dal motore di ricerca Google. Chiunque, con una semplice ricerca, poteva leggere intere chat contenenti dati personali e legali che avrebbero dovuto restare privati.
- Esposizione email in Microsoft Copilot (febbraio 2026): Microsoft ha confermato un errore per cui Copilot ha avuto accesso e ha riassunto email riservate dalle cartelle "Bozze" e "Inviata" senza autorizzazione esplicita, evidenziando i rischi dell'integrazione dell'IA negli spazi di lavoro aziendali.
- Link pubblici su Archive.org: Un ricercatore ha trovato oltre 143.000 conversazioni di Claude, Copilot e ChatGPT accessibili pubblicamente su Archive.org. Si trattava di link "condivisi" che gli utenti credevano privati, ma che erano rintracciabili tramite API pubbliche.

Requisiti: protezione dei dati (2)

Alcuni casi di data leak (involontari?) causati dai maggiori provider AI

- Il caso Amazon: L'azienda ha ufficialmente diffidato i dipendenti dal condividere codice con ChatGPT dopo aver riscontrato casi in cui le risposte del chatbot ricalcavano in modo estremamente fedele dati interni e codice proprietario di Amazon.
- WotNot PII Leak: La startup indiana WotNot ha subito l'esposizione di un database contenente passaporti, cartelle cliniche e curricula caricati dagli utenti. Tali incidenti evidenziano come la mancanza di una sanificazione rigorosa dei dati possa far finire informazioni private nei processi di inferenza o addestramento dei modelli.
- Memorizzazione dei dati di addestramento (GPT-2 e successivi): Ricercatori hanno dimostrato che i modelli linguistici (come GPT-2) possono "rigurgitare" dati di addestramento specifici, inclusi contatti personali di individui che avevano caricato tali informazioni sul web. Questo fenomeno di memorizzazione conferma che ciò che viene inserito durante l'uso può finire per essere replicato quasi testualmente dalle versioni aggiornate del software.
- Integrazione dei dati dei minori: È emerso che diverse aziende, tra cui Google, hanno aggiornato le proprie policy per addestrare i modelli su dati provenienti da adolescenti (previo consenso), sollevando preoccupazioni su come le conversazioni di utenti giovani possano influenzare il comportamento degli agenti AI nelle versioni future.
- …

Requisiti: protezione dei dati (3)

Ait / SCIENZA / Fisica e Matematica

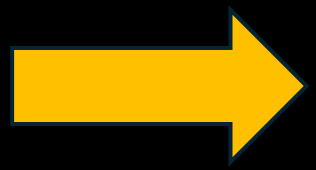
I matematici accusano OpenAI di 'comportamento predatorio'

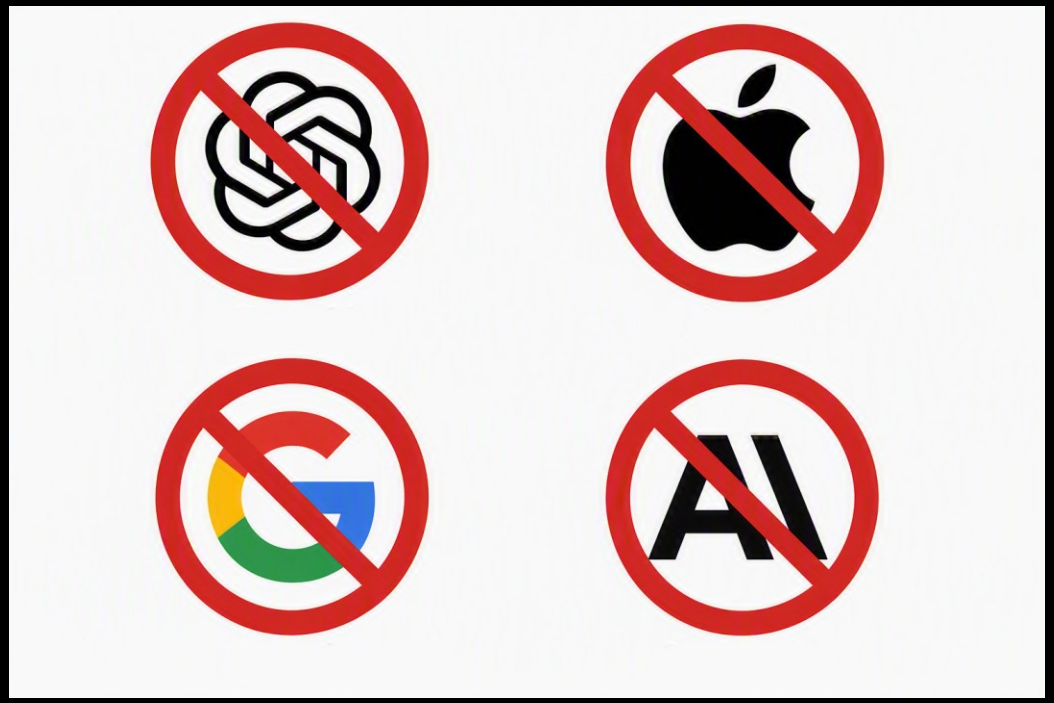
Per 25 vincitori della Medaglia Fields l'IA è una minaccia per il lavoro intellettuale

matematici scendono in campo contro un certo uso dell'**intelligenza artificiale**, dopo che un modello di IA di OpenAI ha risolto uno dei 'problemi del millennio': i ricercatori lanciano l'accusa di '**comportamento predatorio**' perché il modello avrebbe trovato la soluzione **sfruttando illegalmente** le **conversazioni con l'IA di due matematici**.

Requisiti: protezione dei dati (4)







Requisiti: protezione dei dati (5)

Si opera in un contesto *zero-trust*

Come tradurre questa *trustlessness* in requisiti tecnici?

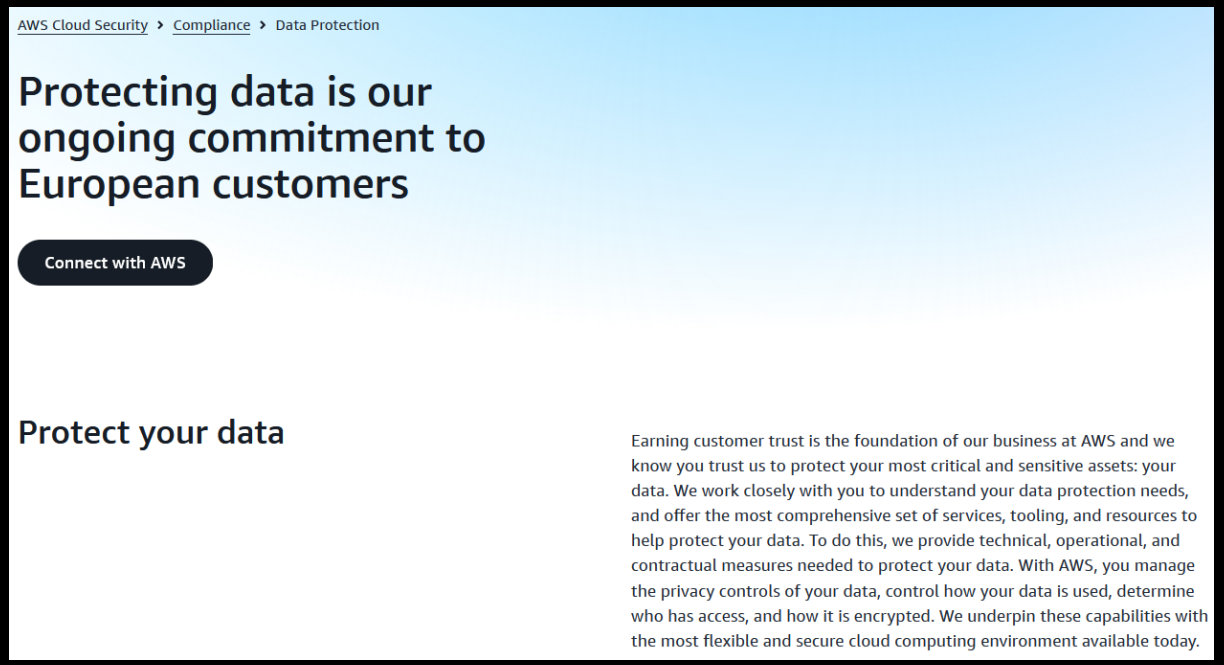
In Retriva, si è ragionato su due livelli:

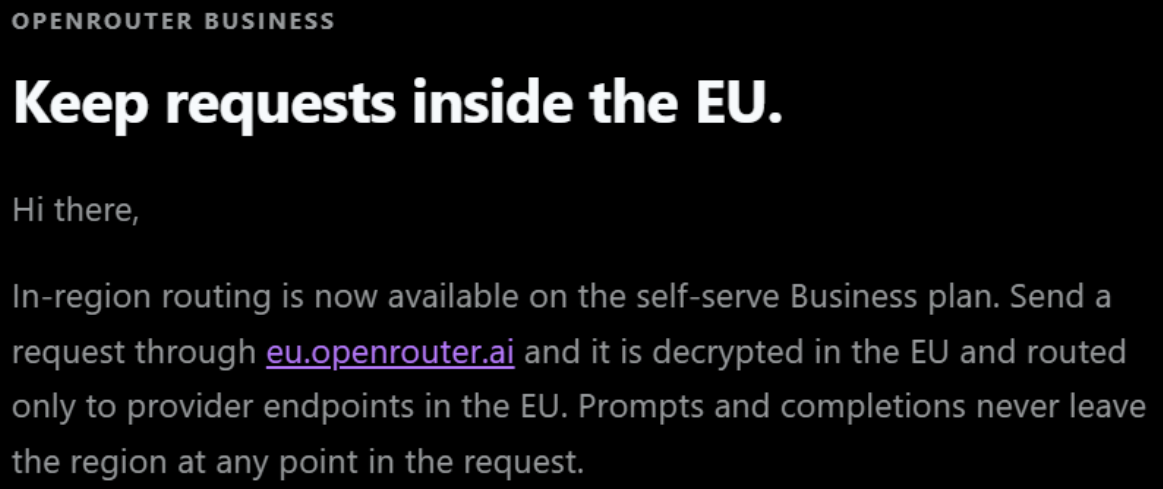
- A livello di sistema: come vedremo tra poco, l'architettura modulare consente di adottare approcci diversi anche il relazione ad una corretta gestione della protezione dei dati (che, come tutti i problemi di security, va modulata in funzione del valore dell'asset che si vuole proteggere).
- Livello feature: nulla vieta di implementare ad un livello più basso altre feature di sicurezza per effettuare un ulteriore irrobustimento del sistema (*hardening*). Esempio: encryption dei dati at rest.

Requisiti: protezione dei dati (6)

Come sta reagendo il mercato?

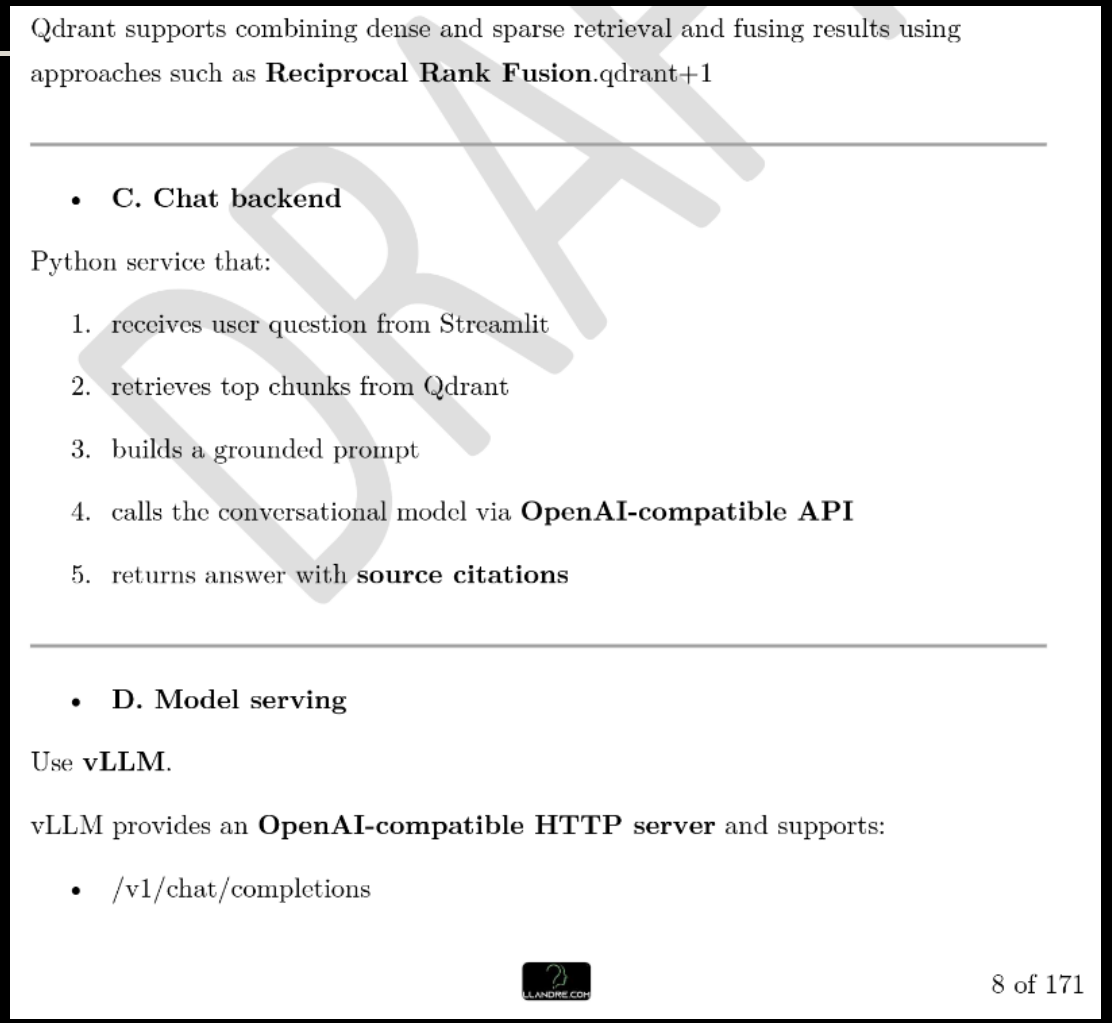
- A conferma di quanto sia sentito il problema, nel mercato stanno comparendo svariate opzioni che soddisfano, in maniera più o meno piena, questa esigenza.
- Commercialmente, si parla di «private AI», «data sovereignty», ecc.





Design

- La fase implementativa è stata preceduta da lunghe sessioni di design, portate avanti con l'ausilio di Microsoft Copilot M365
- Per un'attività di questo tipo, ho selezionato il modello più avanzato al momento disponibile, tipicamente GPT 5.x Think
- Questo tipo di attività può anche essere visto come una efficace forma di *learning by doing* in quanto offre anche numerosi spunti di arricchimento professionale
- Esempio: https://github.com/am-dev-75/retriva-docs/blob/main/design_sessions/Retriva-design_1_0.9.4.pdf



Qdrant supports combining dense and sparse retrieval and fusing results using approaches such as **Reciprocal Rank Fusion**.qdrant+1

- **C. Chat backend**

Python service that:

1. receives user question from Streamlit
2. retrieves top chunks from Qdrant
3. builds a grounded prompt
4. calls the conversational model via **OpenAI-compatible API**
5. returns answer with **source citations**

- **D. Model serving**

Use **vLLM**.

vLLM provides an **OpenAI-compatible HTTP server** and supports:

- /v1/chat/completions

8 of 171

Implementazione

Tooling

Quali tool usare tra i tanti disponibili?

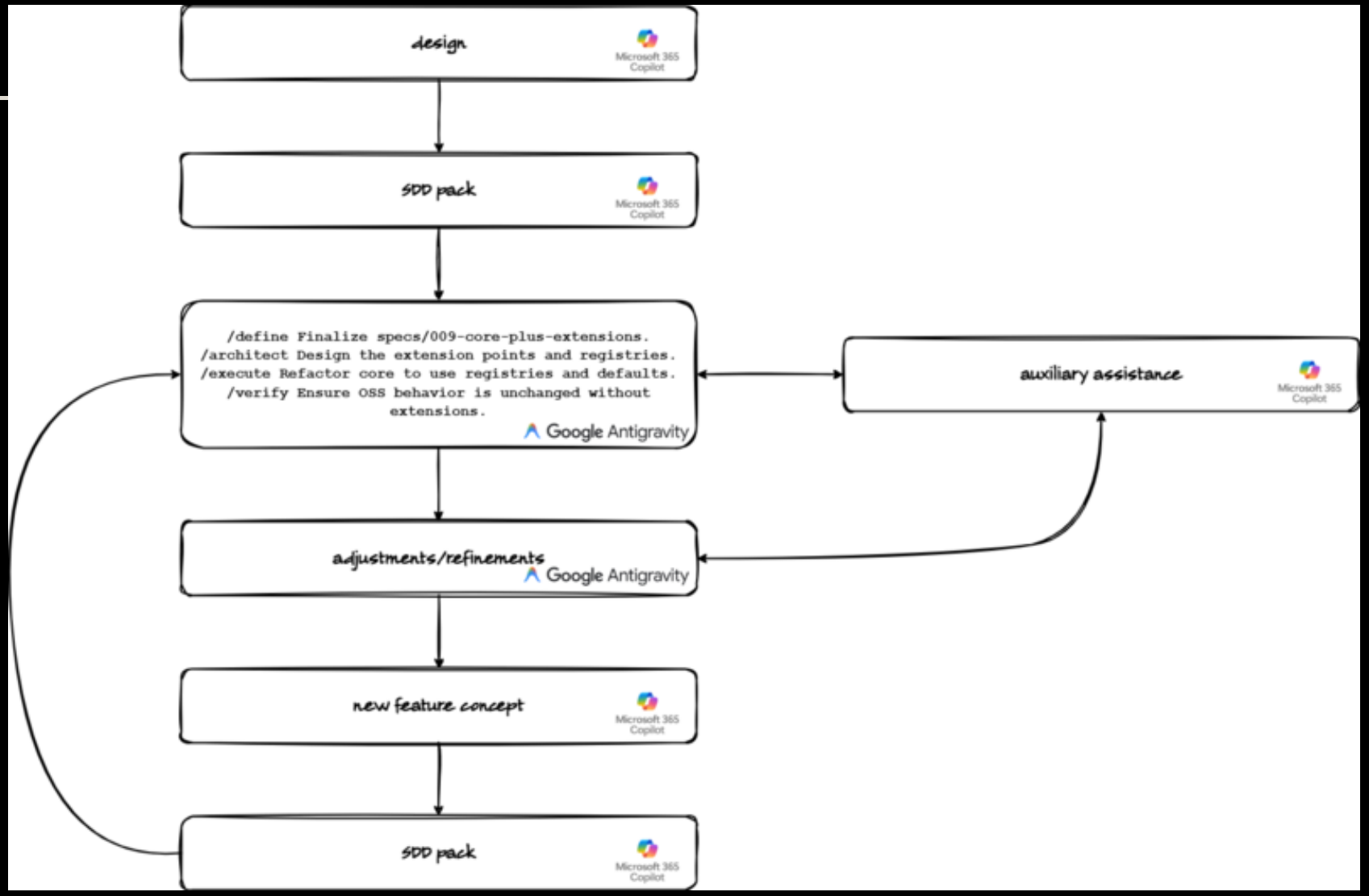
- Microsoft Copilot M365
- ~~Google Antigravity~~
- Microsoft Visual Studio Code (VSCode) combinato con differenti estensioni e LLM
  - ~~Cline extension~~
  - ~~Continue extension~~
  - ~~GitHub Copilot Chat + OpenRouter~~
  - Kilo Code+ OpenRouter
  - LLM
    - GLM 5.2/5.3 Flash
    - DeepSeek v4
    - Opus 4.x
    - ...

Approccio SDD

Testato inizialmente con Google Antigravity, poi abbandonato

Flusso tipico

- `/define`
- `/architect`
- `/execute`
- `/verify`

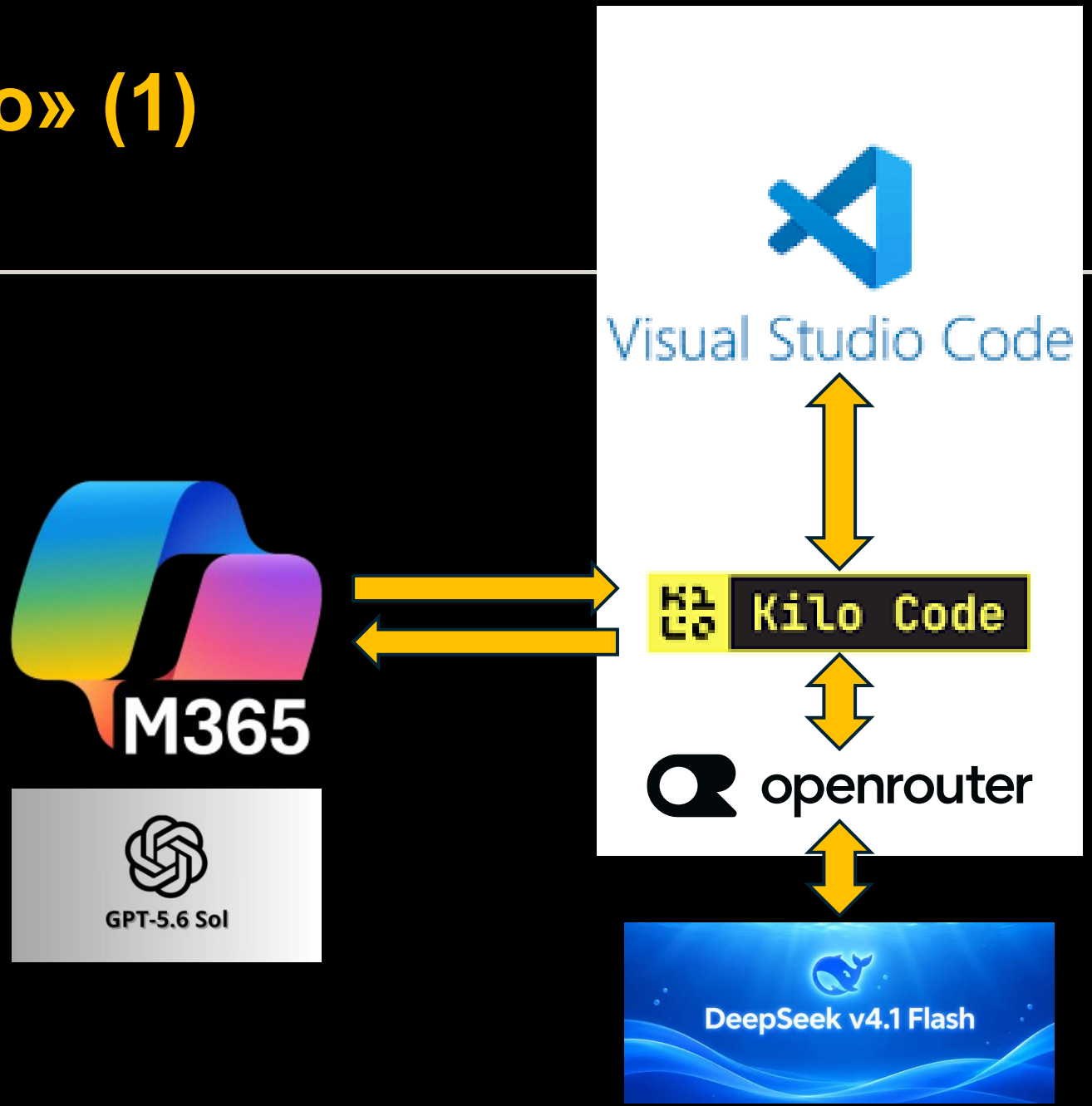


Set up «definitivo» (1)

Architettura dei tool

Articolato in due livelli:

- **Design level**
  - Copilot M365 (GPT 5.x Think)
  - Decisioni sul design a livello architetturale
  - Definizione delle specifiche (sotto forma di prompt)
  - Verifica dei report prodotti dal coding agent
- **Implementation level**
  - Coding effettivo
  - Completa libertà nello scegliere l'LLM più efficiente per lo scopo



Set up «definitivo» (2)

Sono atterrato su un set up che è concettualmente lo stesso di quello suggerito da P. Kiely (Baseten), D. Jackson (Artificial Analysis).

Low-cost coding with open models

CODEX AND CLAUDE CODE

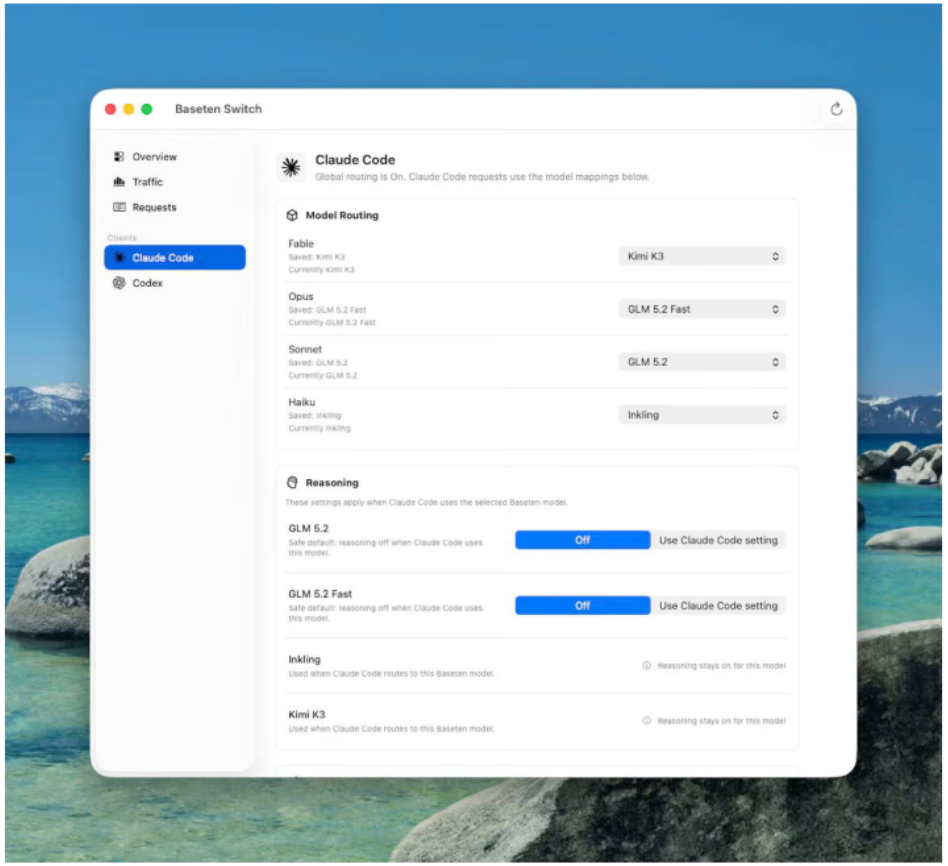
Use Baseten Switch to re-route calls from expensive closed models to low-cost open models without changing your harness

INDEPENDENT HARNESSES

A large number of excellent services like Cursor, Factory, Opencode, and dozens more offer direct access to open models or bring your own key options

TWO MODELS FOR BETTER OUTCOMES

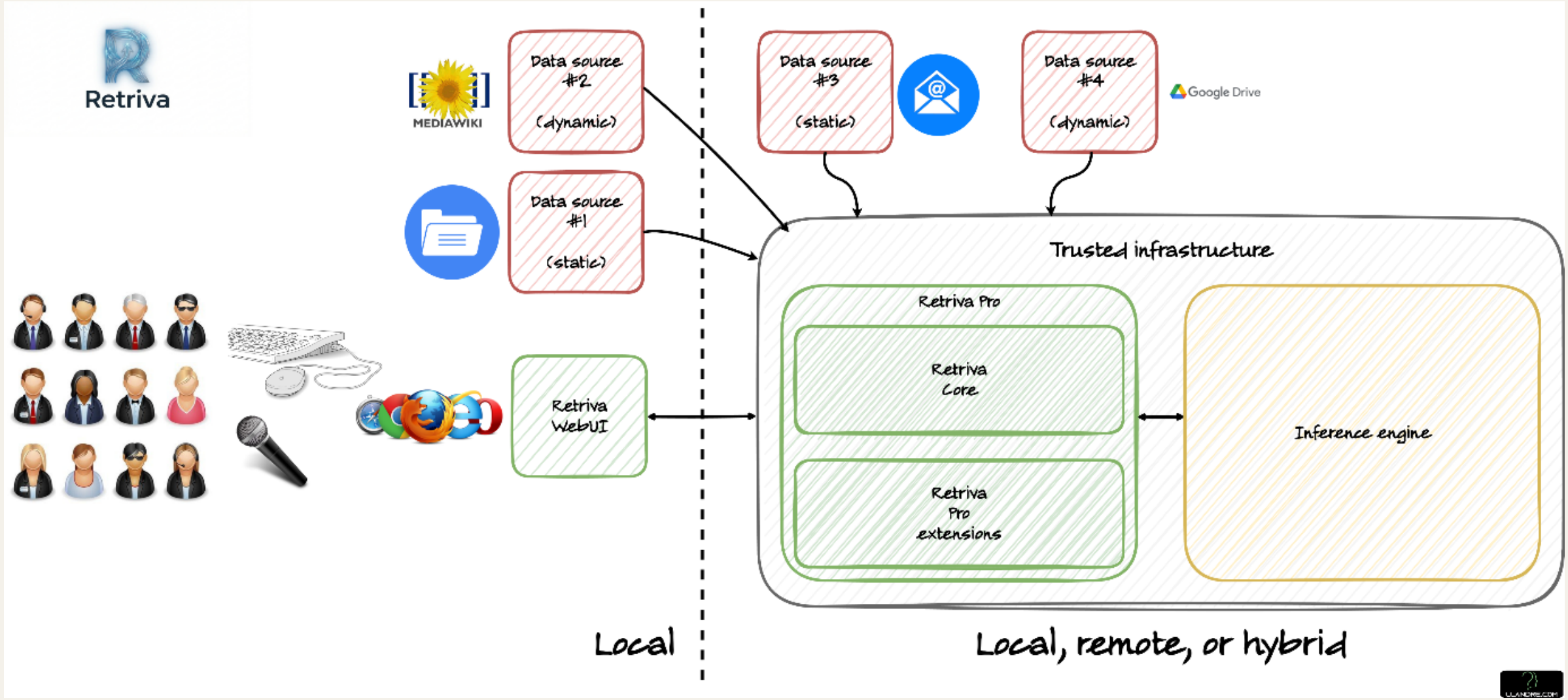
Use GLM-5.3 for long-horizon tasks and main agent work management, let GLM-5.3 Flash handle simple subagent tasks at a lower cost.



Set up «definitivo» (3)

Generazione dei prompt

- Prompt «complessi»: Il prompt da dare in pasto al coding agent viene prodotto da Copilot M365, tipicamente a valle di una sessione di design
  - Il task principale viene automaticamente scomposto in subtask del tutto simili al workflow canonico di Google Antigravity
  - Adatto a implementazioni corpose, come ad esempio la realizzazione di una nuova feature
- Prompt «semplici»: Scritti manualmente direttamente nella chat del coding agent
  - Uso tipico: attività strettamente legate alla codebase su file system locale (**alla quale che Copilot M365 non può accedere**) come bug fix, richieste di chiarimento su alcuni aspetti d'uso, arricchimento della documentazione, ecc.
- Coding agent: VSCode + extension (GitHub Copilot, Cline, Continue, …) + LLM «a scelta» in funzione della complessità del task (*cost optimization*)


Retriva
MEDIAWIKI
Data source #2 (dynamic)
Data source #3 (static)
Data source #4 (dynamic)
Google Drive
Data source #1 (static)
Trusted infrastructure
Retriva Pro
Retriva Core
Retriva Pro extensions
Inference engine
Retriva WebUI
Local
Local, remote, or hybrid

Quasi determinismo

Esempio: stessa domanda posta in due momenti diversi

Q: Elenca le avvertenze di sicurezza nel maneggiare la fusing unit.

A: In base alla documentazione fornita, le avvertenze di sicurezza e le precauzioni operative per il maneggio della fusing unit sono le seguenti:

- **Rischio termico e ustioni:** La fusing unit può raggiungere temperature estremamente elevate. Anche le immediate vicinanze del componente si surriscaldano, pertanto il contatto diretto comporta un elevato rischio di ustioni [1][2].
- **Gestione del peso e rischio di caduta:** Durante la rimozione, è necessario afferrare saldamente i punti di presa specifici indicati nell'illustrazione tecnica (punti [3]). Una presa non corretta o insufficiente potrebbe causare la caduta del componente a causa del suo peso [3].
- **Divieto di smontaggio e modifica:** Le parti interne della fusing unit e la posizione della fusing belt sono calibrate e regolate in fase di produzione prima della spedizione. È espressamente vietato rimuovere viti o procedere a qualsiasi operazione che comporti lo smontaggio dell'unità [4].

Nota: Le indicazioni si riferiscono esclusivamente alle procedure di maneggio e sostituzione manuale riportate nel manuale di servizio. Per interventi di manutenzione interna o regolazioni meccaniche, fare riferimento alle sezioni dedicate del manuale.

Sources:
[1] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — WARNING
[2] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — CAUTION
[3] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — NOTE
[4] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — 2.3.1 Reason for prohibition

Q: Elenca le avvertenze di sicurezza nel maneggiare la fusing unit.

A: In base alla documentazione tecnica fornita, le avvertenze di sicurezza per la maneggiatura della fusing unit sono le seguenti:

- **Rischio termico e ustioni:** Il componente può raggiungere temperature estremamente elevate [1]. Anche le immediate vicinanze si surriscaldano, pertanto il contatto diretto con l'area può causare ustioni [2].
- **Gestione del peso e rischio di caduta:** Durante la rimozione, è necessario afferrare saldamente i punti di presa indicati come [3] nell'illustrazione di riferimento, assicurandosi di sostenere adeguatamente il peso del modulo per prevenirne la caduta [3].
- **Divieto di smontaggio:** Le parti interne e l'allineamento della fusing belt sono tarati in fase di produzione prima della spedizione. È severamente vietato rimuovere qualsiasi vite che possa portare allo smontaggio dell'unità [4].

Sources:
[1] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — WARNING
[2] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — CAUTION
[3] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — NOTE
[4] bizhub C360i_C300i_C250i_C361i_C301i_C251i_SM_EN_20231227(1).pdf — 2.3.1 Reason for prohibition

Set flessibile di LLM per l’inferenza a run-time

Sono sufficienti poche variabili per cambiare il set di modelli utilizzati

```
# --- LLM / embedding / vision provider (OpenAI-compatible) --------------------
OPENAI_PROVIDER_API_KEY=replace-me

# Single knob to redirect ALL traffic to a different OpenAI-compatible
# provider (e.g. a local vLLM/Ollama server). Must include the /v1 suffix.
# …
LLM_PROVIDER_BASE_URL=https://openrouter.ai/api/v1

# Optional PER-PURPOSE base-URL overrides. Leave commented to inherit
# …
#EMBEDDING_BASE_URL=https://openrouter.ai/api/v1
#VISUAL_BASE_URL=https://openrouter.ai/api/v1
#CHAT_BASE_URL=https://openrouter.ai/api/v1
#RETRIEVAL_RERANK_BASE_URL=https://openrouter.ai/api/v1

# Optional PER-PURPOSE API-key overrides. Leave commented to inherit
# …
#EMBEDDING_API_KEY=
#VISUAL_API_KEY=
#CHAT_API_KEY=
#RETRIEVAL_RERANK_API_KEY=

CHAT_MODEL=qwen/qwen3.6-27b
EMBEDDING_MODEL=baai/bge-m3
VISUAL_MODEL=qwen/qwen3-vl-32b-instruct
RETRIEVAL_RERANK_MODEL=cohere/rerank-v3.5
```

LOCs (al 6 Ottobre 2026)

Scritte a mano: 0

```
amarson@LW11Y24037:/mnt/devel/retriva$ manifest=$(mktemp)
…
cloc --list-file="$manifest«
…
```

Language	files	blank	comment	code
Python	629	24785	36358	137869
Markdown	394	8955	150	39637
YAML	24	802	440	15007
JSON	32	4	0	9083
TypeScript	46	668	911	5957
CSS	19	673	317	3740
SQL	25	361	872	3448
Bourne Shell	9	93	123	746
Text	10	4	0	295
Dockerfile	6	88	91	192
TOML	6	22	39	141
INI	1	10	0	44
JavaScript	1	2	15	27
Mako	1	8	0	17
HTML	1	1	15	13
SVG	2	0	0	2
SUM:	1206	36476	39331	216218

Protezione dei dati riservati

Ambito estremamente fluido

Premessa importante: visto il campo di applicazione per cui è pensato Retriva, naturalmente si esclude la possibilità di utilizzare di provider come Anthropic, Microsoft, OpenAI, ecc. Al di là delle garanzie formali che dichiarano, i loro server sono comunque sottoposti alla legge USA (CLOUD Act). Fissato questo, quali soluzioni sono praticabili?

I trend principali che mi sembra di poter indicare sono:

- Deployment locale (inferenza compresa)
- Su cloud europeo (pubblico e/o privato)
- Approccio ibrido (parte on-prem, parte in cloud)

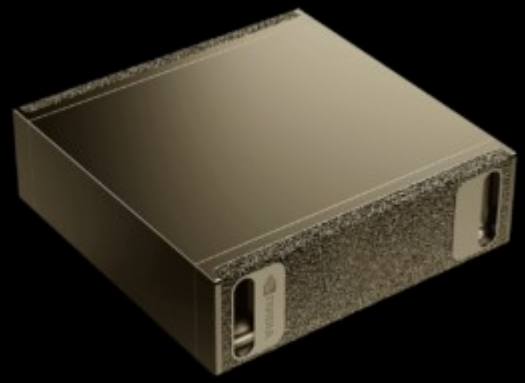
Approcci per il deployment

Deployment	Security	Capex	Opex
Ibrido/cloud	Livello “**standard**”: provider europei e server in EU in modo che questi non sottostiano allo US CLOUD Act. L’«in-region routing» di OpenRouter va esattamente in questa direzione.	$/$$	$/$$
Ibrido/cloud	Livello “**dati biomedici**”: ai requisiti del livello precedente, si aggiungono ulteriori certificazioni specifiche (vedi ad esempio https://www.ovhcloud.com/it/solutions/industries/manufacturing/).	Varia in funzione di quanto si ripartisce l’infrastruttura tra on-prem e cloud.	Varia in funzione di quanto si ripartisce l’infrastruttura tra on-prem e cloud.
Ibrido/cloud	Livello “**finanza**”: la parte in cloud è *confidential computing enabled*.	Varia in funzione di quanto si ripartisce l’infrastruttura tra on-prem e cloud.	Varia in funzione di quanto si ripartisce l’infrastruttura tra on-prem e cloud.
On-prem	Può arrivare al **military-grade**.	$$$	$

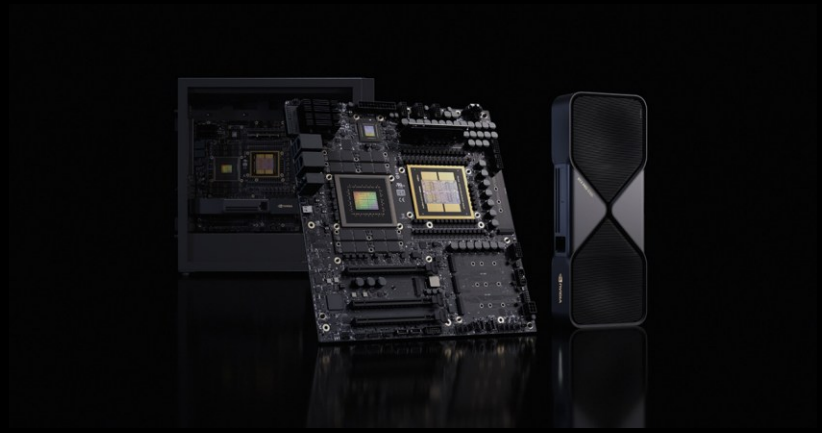
Approcci per il deployment: on-prem

- Segmento di mercato estremamente frizzante
- Probabilmente un giorno si arriverà ad avere un rapporto costo hw/potenza di calcolo tale da rendere accessibile questo approccio a tutte le SME
- In ogni caso **non va assunto che automagicamente garantisca un livello di sicurezza superiore alle altre soluzioni**
  - Un deployment on-prem mal configurato potrebbe essere più vulnerabile di uno realizzato in cloud
- Costi capex: a parte l'acquisto dell'hardware, non vanno trascurati i costi NRE di deployment dei modelli (per chi l'ha vissuta, ricorda la «battaglia» con la limitatezza delle risorse dei microcontrollori)
  - Formati ottimizzati
  - Quantizzazione
  - ...
- Rischio obsolescenza

NVIDIA DGX Sparc (~5.000$)



NVIDIA DGX Station (~100.000$)



Approcci per il deployment: cloud

- Come riassunto nella tabella precedente, si possono ottenere livelli di protezione più o meno elevati
- Come in tutte le questioni di sicurezza, si tratta di **stabilire l'altezza dell'asticella in maniera commisurata al valore degli asset da proteggere**

Confidential computing

- OVH lo definisce come

  Il Confidential Computing, in italiano anche "elaborazione confidenziale" o "informatica confidenziale", è una tecnologia per la protezione continua dei dati durante l'elaborazione, essenziale in specifici ambiti e utilizzi.

- Tipicamente sfrutta un tecnologia hardware che fa in modo che il contenuto della RAM (sia quella della CPU, sia quella della GPU) sia crittografato
  - Siccome la chiave crittografica non è accessibile al provider cloud, questo non può accedere ai dati che transitano (query, risposte, embedding, chunk, ecc.)
  - Abilitando la crittografia anche per i dati at rest, si rende completamente inaccessibile il contenuto della macchina al provider
- Sicuro al 100%? No! ⚠️
  - https://arxiv.org/abs/2404.03526
  - https://www.amd.com/en/resources/product-security/bulletin/amd-sb-3020.html
  - https://www.amd.com/en/resources/product-security/bulletin/amd-sb-3014.html
  - ...

Quale trade-off scegliere per Retriva?

- Per quanto riguarda il caso d'uso tipico di Retriva, assumiamo che sia sensato puntare a livello «finanza» in fatto di protezione dei dati
  - Di conseguenza, **imponiamo come requisito che, per le eventuali porzioni del sistema deployate in cloud, sia abilitato il *confidential computing***
- Per motivi di costo, assumiamo che una soluzione fully on-prem non sia praticabile
  - Essendo ancora in una fase sperimentale, i benefici di Retriva in azienda non sono quantificabili, pertanto il management non può fare una valutazione costi/benefici ⇒ preferibile in questo momento minimizzare i costi fissi, accettando dei costi variabili, anche relativamente alti, con una logica *pay-as-you-go*
  - Tenendo anche conto del limitato budget temporale, non c'è materialmente la possibilità investire energie in *inference engineering*, concentrando invece gli sforzi sugli aspetti architetturali e funzionali

Deployment ibrido

Esempio: deployment prototipale in azienda

Contesto

- On-prem
  - Servizi core
- In cloud con *confidential computing*
  - LLM runner
- Retriva è agnostico rispetto ai server LLM
  - Nel momento in cui questi espongono un'API compatibile con OpenAI, è sufficiente cambiare alcuni setting nel file di configurazione per interfacciare Retriva
- Nelle slide seguenti, maggiori dettagli sul deployment del modello utilizzato per l'embedding dei chunk (BAAI/bge-m3)

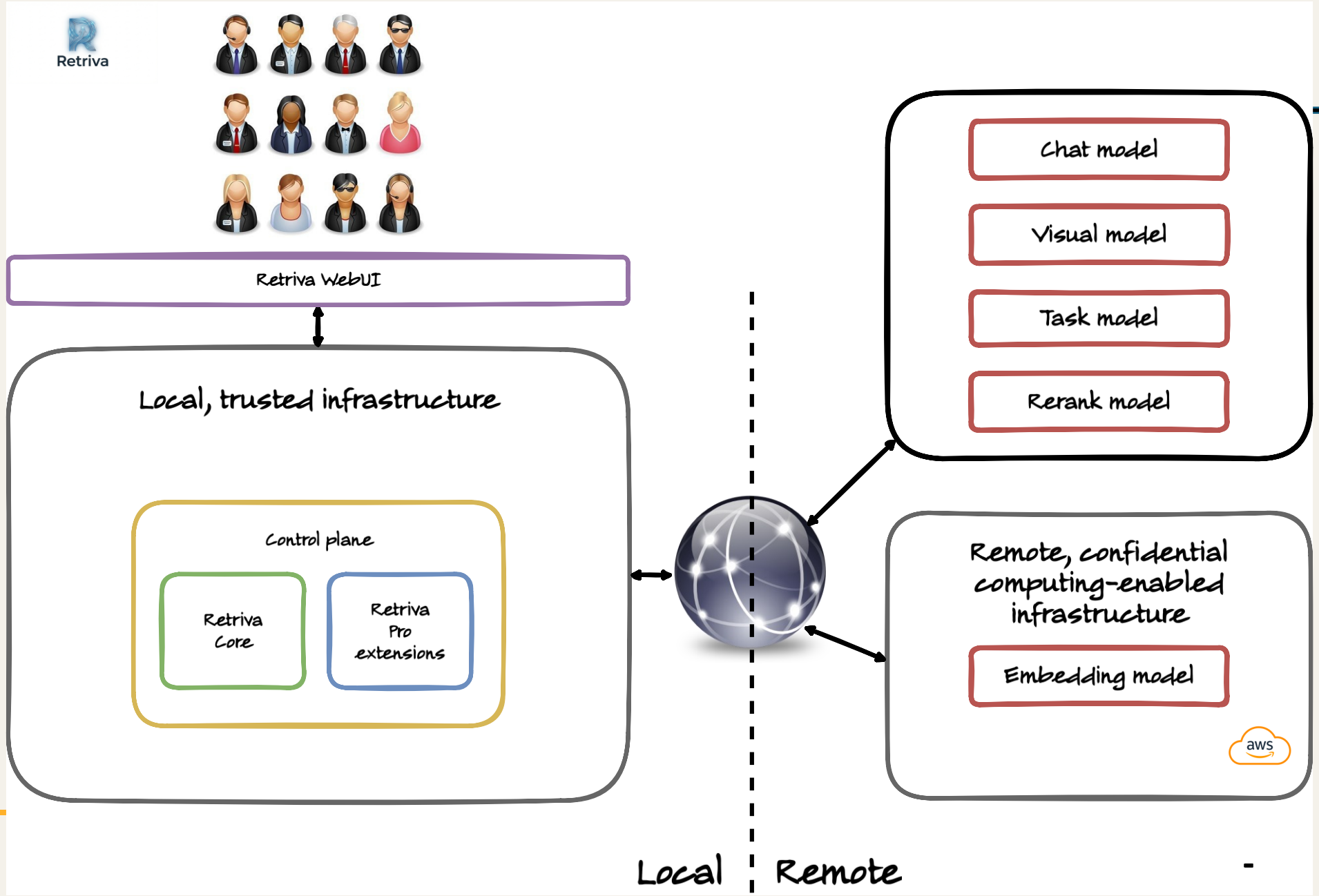
Tooling e tecnologie

- Mi contraddico (consapevolmente): per semplicità, nell'esempio si utilizza AWS (quindi soggetto al CLOUD) come cloud provider; i concetti esposti rimangono validi
- Inevitabilmente, la parte nascosta dall'API OpenAI è provider dipendente
- AMD SEV-SNP (Secure Encrypted Virtualization-Secure Nested Paging)
- Implementazione e deployment realizzati con Copilot M365








Retriva
Retriva WebUI
Local, trusted infrastructure
Control plane
Retriva Core
Retriva Pro extensions
Chat model
Visual model
Task model
Rerank model
Remote, confidential computing-enabled infrastructure
Embedding model
aws
Local
Remote

Deployment

Creazione della VM

```
amarson@LW11Y24037:/mnt/devel/retriva/retriva-confidential-poc$ ./01-create-sev-snp-vm.sh
[2026-08-03T09:42:27Z] Checking AWS credentials
[2026-08-03T09:42:28Z] AWS account: …
[2026-08-03T09:42:28Z] Caller: arn:aws:iam::472670657019:user/amarson
[2026-08-03T09:42:28Z] Region: eu-west-1
[2026-08-03T09:42:28Z] Checking AMD SEV-SNP support for m6a.2xlarge in eu-west-1
[2026-08-03T09:42:29Z] m6a.2xlarge reports AMD SEV-SNP support
[2026-08-03T09:42:29Z] Selecting an available Availability Zone
[2026-08-03T09:42:29Z] Selected Availability Zone: eu-west-1a
[2026-08-03T09:42:29Z] Finding the latest Canonical Ubuntu Server 24.04 LTS x86_64 AMI
…
```

Deployment

Verifica abilitazione SEV-SNP (1)

```
amarson@LW11Y24037:/mnt/devel/retriva$ aws ec2 describe-instance-types \
  --instance-types m6a.2xlarge \
  --query 'InstanceTypes[0].{
    InstanceType:InstanceType,
    SupportedFeatures:ProcessorInfo.SupportedFeatures,
    VCpuCount:VCpuInfo.DefaultVCpus,
    MemoryMiB:MemoryInfo.SizeInMiB
  }' \
  --output json
{
    "UserId": «…",
    "Account": «…",
    "Arn": "arn:aws:iam::…:user/…"
}
{
    "InstanceType": "m6a.2xlarge",
    "SupportedFeatures": [
        "amd-sev-snp"
    ],
    "VCpuCount": 8,
    "MemoryMiB": 32768
}
```

Deployment

Verifica abilitazione SEV-SNP (2)

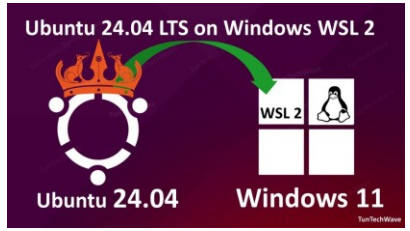
```
amarson@LW11Y24037:/mnt/devel/retriva/retriva-confidential-poc$ aws \
  --profile "$AWS_PROFILE" \
  --region "$AWS_REGION" \
  ec2 describe-instances \
  --instance-ids "$INSTANCE_ID" \
  --query 'Reservations[0].Instances[0].CpuOptions' \
  --output json
{
    "CoreCount": 4,
    "ThreadsPerCore": 2,
    "AmdSevSnp": "enabled"
}
```

Deployment

Verifica abilitazione encryption dello storage non volatile

```
amarson@LW11Y24037:/mnt/devel/retriva/retriva-confidential-poc$ aws \
  --profile "$AWS_PROFILE" \
  --region "$AWS_REGION" \
  ec2 describe-volumes \
  --volume-ids "$ROOT_VOLUME_ID" \
  --query 'Volumes[0].{
    VolumeId:VolumeId,
    Encrypted:Encrypted,
    Type:VolumeType,
    SizeGiB:Size,
    KmsKeyId:KmsKeyId,
    State:State
  }' \
  --output table
```

DescribeVolumes					
Encrypted	KmsKeyId	SizeGiB	State	Type	VolumeId
True	arn:aws:kms:eu-west-1:…:key/b398e068-9f03-4f61-a53e-c04aa6887756	60	in-use	gp3	vol-0784f6e365a91e17b



retriva-core

retriva-gateway

retriva-webui

retriva-crm-assistant

retriva-mediawiki-connector

retriva-qdrant

retriva-ingestion

retriva-tika

retriva-messaging-extension

retriva-searxng

retriva-redis

retriva-worker

retriva-whisper

retriva-email-agent-connector

docker engine

relay

SSM

Chat model

Rerank model

Visual model

Task model

openrouter

SSM Agent

Text Embeddings Inference (TEI)

Embedding model

aws

Canale di comunicazione

- SSM: AWS Systems Manager (Systems Manager) was formerly known as "Amazon Simple Systems Manager (SSM)" and "Amazon EC2 Systems Manager (SSM)".
- SSM Agent in esecuzione nella macchina remota (evita l'utilizzo della porta SSH aperta)






```
Docker container
      ↓
host.docker.internal:18081
WSL relay 0.0.0.0:18081
      ↓
WSL SSM tunnel
127.0.0.1:18080
      ↓
AWS VM 127.0.0.1:8080
      ↓
BGE-M3
```

Analisi della soluzione (1)

Quanto è sicura la nostra implementazione?

- Enabling SEV-SNP activates the hardware protections.
- Attestation provides cryptographic evidence that those protections and a particular initial VM state are present.
- `"AmdSevSnp": "enabled"` ⇒ That demonstrates that AWS launched the instance with the SEV-SNP option. It is an AWS control-plane assertion, however—**not independent cryptographic proof**.

The basic attestation problem is:

A client is about to send sensitive data to a remote machine. How can it know that the remote machine is a genuine confidential VM, configured according to an acceptable security policy, rather than an ordinary VM or an impostor?

The VM answers by requesting an **attestation report from the AMD Secure Processor**. The report contains security-relevant claims and a cryptographic signature.

Analisi della soluzione (2)

Quanto è sicura la nostra implementazione?

The remote verifier then asks:

1. Is this a structurally valid SEV-SNP report?
2. Was it signed by a key endorsed by AMD?
3. Does that endorsement chain to an AMD root of trust?
4. Does the report describe an acceptable platform and TCB?
5. Does its launch measurement match something I trust?
6. Is the report fresh rather than replayed?
7. Is it bound to the server connection I am about to use?
8. Is it bound to the application and model that I intended to contact?

Analisi della soluzione (3)

Quanto è sicura la nostra implementazione?

Who signs the report on AWS?

Your EC2 instance uses shared tenancy. AWS therefore uses a VLEK, or Versioned Loaded Endorsement Key, for report signing. AWS distinguishes:

- **VCEK** for applicable Dedicated Host scenarios;
- **VLEK** for shared-tenancy instances.

Both ultimately rely on an AMD certificate chain, but the key provisioning and identity semantics differ.

Analisi della soluzione (4)

Quanto è sicura la nostra implementazione?

The trust chain is approximately:

```
AMD root certificate
↓ signs
AMD VLEK intermediate/signing hierarchy
↓ endorses
AWS-hosted VLEK certificate
↓ verifies signature on
SEV-SNP attestation report
```

Analisi della soluzione (5)

Quanto è sicura la nostra implementazione?

What your current result actually proves

Your current evidence supports this claim:

An EC2 VM configured for AMD SEV-SNP successfully requested an AMD SEV-SNP attestation report. Its signature was validated with a VLEK certificate, and that VLEK certificate was validated against AMD’s Milan certificate chain. The report’s TCB information was consistent with the VLEK certificate.

That is meaningful hardware-backed evidence.

However, …

Analisi della soluzione (6)

Quanto è sicura la nostra implementazione?

… it does **not yet prove** all of the following:

- that the report was requested in response to a nonce generated by Retriva;
- that BAAI/bge-m3 was running when the report was created;
- that the currently running TEI (1) container matches the image digest you intended;
- that the downloaded model files match an approved manifest;
- that port 8080 belongs to the attested workload;
- that the SSM tunnel terminates at the same confidential VM that produced the report;
- that the API key was generated and remains only inside the confidential VM;
- that the current state is unchanged since the report was generated.

(1) Text Embedding Inference

Debolezze (1)

Your current check asks:

Is this a genuine, untampered SEV-SNP report signed by an AMD-endorsed key?

A complete verifier must additionally ask:

Even if genuine, is this particular report acceptable to me?

For example, a genuine report could describe:

- an unexpected launch measurement;
- an outdated TCB;
- an unapproved guest policy;
- an old but previously valid report;
- a different public key;
- a different application build.

Debolezze (2)

Why verifying inside the VM is weaker

Your current script performs verification inside the same VM that produced the report:

```
Confidential VM
├── requests report
├── downloads AMD chain
└── verifies its own report
```

A compromised guest could simply print:

```
vlek.pem: OK

VEK signed the Attestation Report!

...
```

Conclusioni su abilitazione SEV-SNP

- Nonostante l'approccio adottato, si è visto come si possano implementare svariati interventi di hardening per alzare ulteriormente il livello di security. Al solito, sta all'ingegnere di sistema stabilire quanto in alto ha senso porre l'asticella.
- Per semplicità di esposizione, nell'esempio il *confidential computing* è stato abilitato per una piccola parte dell'architettura, sufficiente comunque per illustrare il concetto. È importante sottolineare come **analogo procedimento si può replicare anche per una macchina dotata di GPU Nvidia**, **abilitando quindi l'encryption in hardware anche per la sua VRAM**.

La doppia anima di Retriva: probabilistica e deterministica

Il pericolo dell'overengineering

- La potenza degli LLM può avere un effetto pericoloso, cioè la **tendenza a usarli per risolvere qualsiasi problema anche quando sarebbe opportuno adottare soluzioni tradizionali, meno complesse e intrinsecamente deterministiche**.
  - Attenzione a chi non si fa molti scrupoli a cavalcare questa moda per proprio tornaconto, spingendo i propri clienti a comprare soluzioni overingegnerizzate e sovradimensionate!
- Mi è capitata una situazione di questo tipo nello sviluppo di una delle funzionalità del modulo *CRM Assistant*: stavo per basare l'implementazione sull'utilizzo della KB combinata con la pipeline di retrieval, quando la stessa funzionalità si può ottenere optando per un classico DB relazionale, ottenendo una soluzione più efficiente e più robusta.
- In questo modo, si riescono a coniugare due «anime», quella probabilistica e quella deterministica, facendole lavorare assieme e prendendo il meglio dei due mondi. In pratica, si mantiene la possibilità per l'utente di interagire con il sistema in linguaggio naturale, utilizzando dietro le quinte una solida e collaudatissima base di dati tradizionale.
- Rule of thumb: **Usare software «tradizionale» (ovvero deterministico) se un requisito si può ottenere per questa via**. **Usare l'AI** (ovvero l'approccio probabilistico) **in alternativa o in combinazione con il software classico quando quest'ultimo non permette di raggiungere l'obiettivo (per infattibilità o per costi insostenibili)**.

CRM Assistant: qualifica dei nuovi lead

Esempio della doppia anima di Retriva (probabilistica e deterministica)

Scopo: efficientare il procedimento di qualifica di un nuovo lead, ovvero:

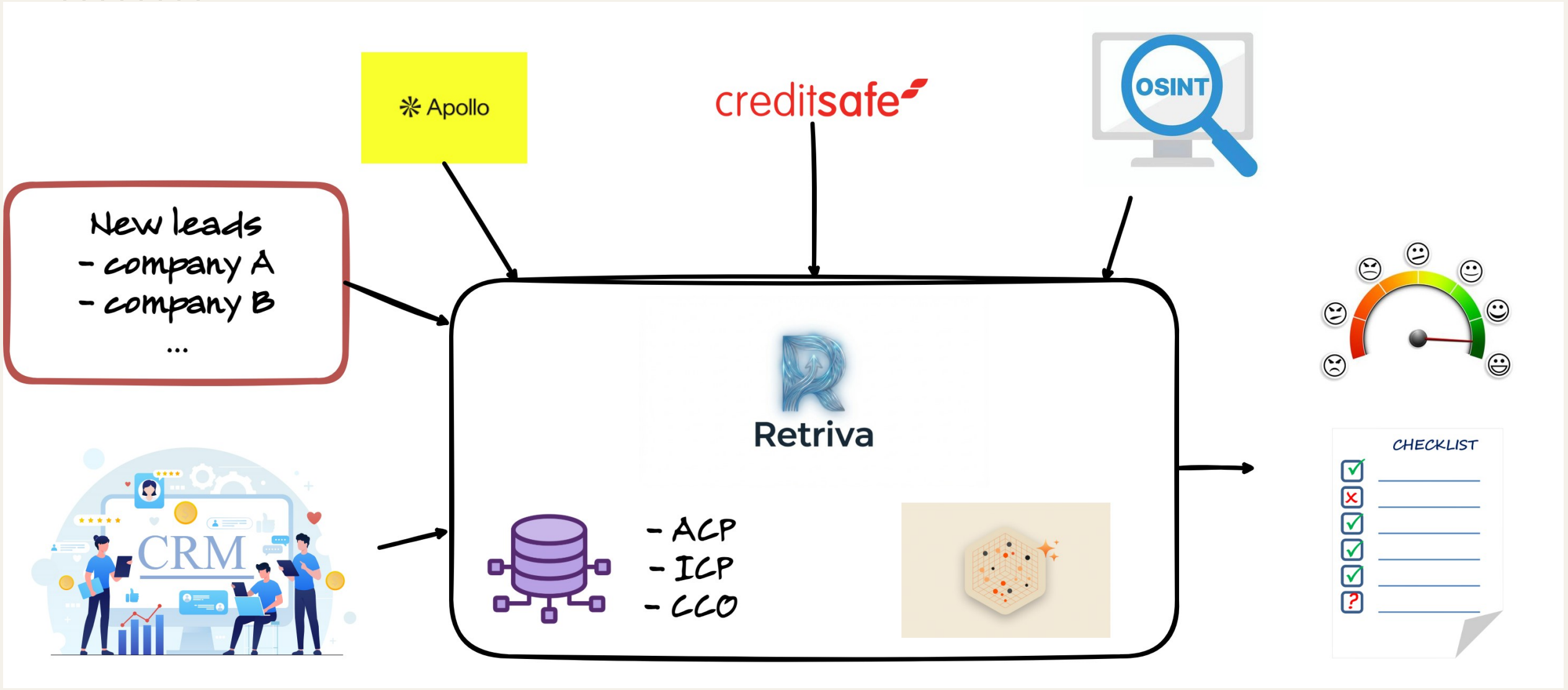
- A partire dall'anagrafica del lead, completarne l'*enrichment* in modo da avere una descrizione più ricca, cioè un profilo più dettagliato
- Esprimerne uno score complessivo che sintetizza quanto questo sia aderente rispetto a:
  - Average Customer Profile (ACP)
    - Prodotto da un LLM sulla base dei clienti già acquisiti (comunque modificabile dal personale umano)
  - Ideal Customer Profile (ICP)
    - Scritto dal personale umano
  - Commercial Company Offering (CCO)
    - Prodotto da un LLM sulla base di ciò che noi offriamo in termini di prodotti e servizi (comunque modificabile dal personale umano)

  E se sia sufficientemente solido finanziariamente.
- In formula:

$$overall_score = (ACP_weight * ACP_score + ICP_weight * ICP_score + CCO_weight * CCO_score) * credit_check(lead)$$

$$credit_check(lead) = I_A(lead.credit_score >= credit_threshold)$$

I_A: Indicator function.


Apollo
creditsafe
OSINT
New leads
- company A
- company B
...
Retriva
CRM
- ACP
- ICP
- CCO
CHECKLIST

Candidate	Candidate	ACP Fit on	ACP Evide	ACP Dime	ACP Dime	Offering Match Status	Grounding Reason	Outreach Readiness	Primary Offering Family (Grounded)	Fa
STM Products S.r.l.	possible_fit	0,576	0,559	2	4	no_grounded_offering	NO_RELEVANT_CCO_CAPABILITY	needs_discovery_call		
JMA SERVIZI E INDUSTRIA S.R.L.	insufficient_evidence	0,105	0,265	1	4	no_grounded_offering	INSUFFICIENT_CANDIDATE_INFORMATION	needs_additional_research		
El.En. S.p.A.	insufficient_evidence	0	0	0	4	no_grounded_offering	NO_DISTINCTIVE_PUBLIC_SIGNALS	needs_additional_research		
Texa Industries	possible_fit	0,601	0,559	2	4	no_grounded_offering	NO_DISTINCTIVE_PUBLIC_SIGNALS	needs_discovery_call		
Società Cooperativa Bilanciai	identity_unresolved					insufficient_offering_evidence		needs_additional_research		
365FarmNet GmbH	insufficient_evidence	0,211	0,265	1	4	grounded_offering_match	SUFFICIENT_DISTINCTIVE_EVIDENCE	needs_additional_research	TOLOMEO Embedded Manager	
3N Kompetenzzentrum Niedersachsen Netzwerk Nachwachsende Rohstoffe e.V.	insufficient_evidence	0	0	0	4	no_grounded_offering	NO_DISTINCTIVE_PUBLIC_SIGNALS	needs_additional_research		
4Disc GmbH	insufficient_evidence	0,158	0,265	1	4	grounded_offering_match	SUFFICIENT_DISTINCTIVE_EVIDENCE	needs_additional_research	TOLOMEO IIoT Ecosystem	
ABO Energy GmbH & Co. KGaA	weak_fit	0,347	0,5	2	4	grounded_offering_match	SUFFICIENT_DISTINCTIVE_EVIDENCE	ready_for_review	Embedded Software Development	

Sviluppi in fase di prototipazione

CRM Assistant: RMA Eligibility Assessment

- ToloMEO: ecosistema per lo sviluppo di applicazioni IIoT (dal dispositivo edge fino alle applicazioni cloud)
- Tra i servizi implementati, c'è un sistema di gestione del ciclo di vita dei dispositivi (Product Management System o PMS)
- PMS è indispensabile per gli addetti al post vendita, in particolare per l'autorizzazione degli RMA in arrivo dai clienti
- Una possibile funzionalità del CRM Assistant di Retriva è l'«RMA Eligibility Assessment», ovvero uno strumento per efficientare il delicato processo di valutazione degli RMA
- Questo è possibile grazie al fatto che nella propria KB Retriva dispone delle informazioni da correlare per prendere la decisione
  - Storia dello specifico s/n a cui si riferisce l'RMA (grazie a PMS)
  - Storia del modello a cui l'esemplare appartiene (grazie all'MRP)
  - Profilo del cliente (grazie all'ERP, al CRM, alle interazioni dirette, ecc.)

TOLOMEO Cloud

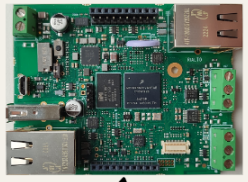
IoT services

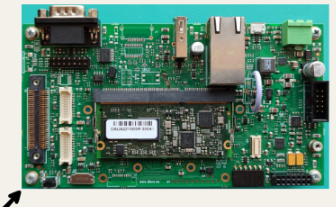
Cybersec (CRA-ready)

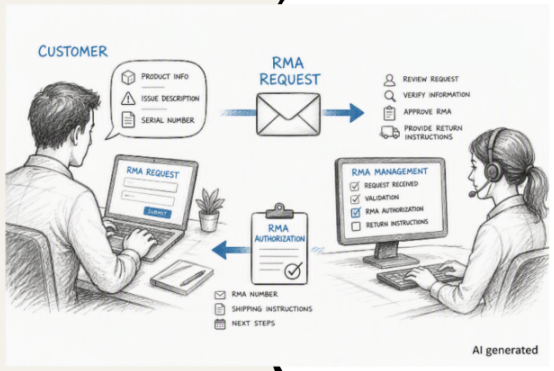
...

PMS









Retriva

ERP/MRP/CRM

DAVE Embedded Systems

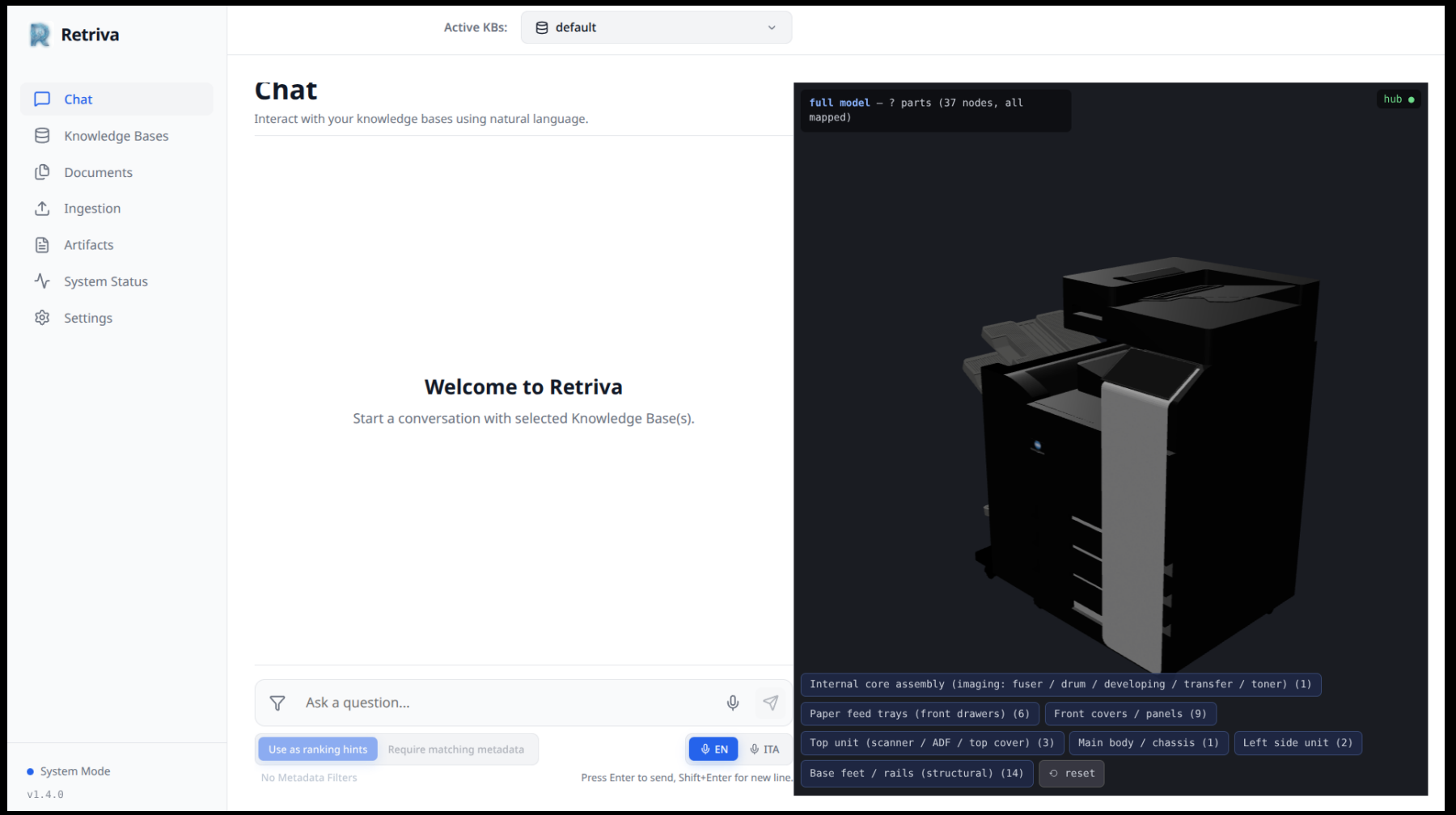
authN/authZ

- Ancora il concetto di modularità ⇒ IAM esterno
- Microsoft Entra per adattarsi al sistema informativo esistente



3D Service Assistant

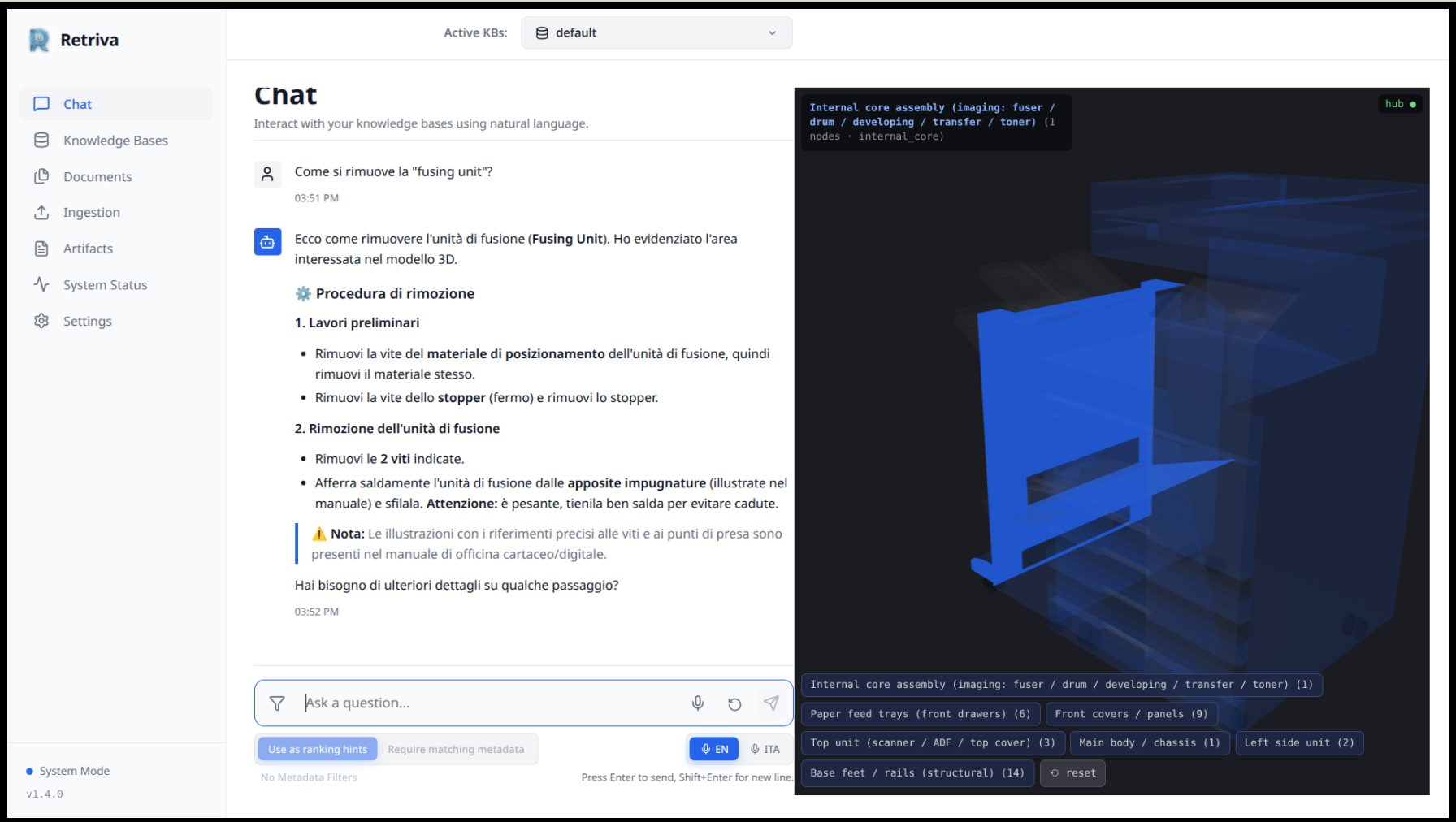
Visualizzatore modello 3D inglobato nella web application



Sviluppato In collaborazione con Ambiens.

3D Service Assistant

Output dell’agente conversazione arricchito e complementato da modello 3D navigabile.



Conclusioni



Salvatore Sanfilippo aka antirez

Lui [David Heinemeier Hansson, creatore di Ruby on Rails, NdR] ha fatto un intervento importante [...] in cui dice sostanzialmente alcune cose che io fatico a capire come possono essere contrastate, cioè dire che oggi per la maggior ***per la grandissima maggioranza dei business scrivere codice a mano o andare a vagliare manualmente il codice scritto dalla macchina sia una follia è la sacrosanta verità****. Io questa cosa la ripeto qua, vi dico in più che penso che un programmatore o una programmatrice d'esperienza che abbiano provato i modelli come Astra, ma anche Fable 5.1, [...] però chi ha avuto modo di accedere a questi modelli di frontiera e usarli mettendosi una mano nella coscienza, ...* ***e non è uscito convinto che la stragrande maggioranza del codice si deve scrivere semplicemente con lei, trasformandoti solo in progettista e non più in programmatore****. Io una persona che ha che avesse tale opinione che quindi lei ancora non è buona per scrivere la maggior parte del codice, io non la assumerei nella mia eh azienda, perché credo che una visione talmente fuori dal mondo, talmente distorta che quella persona finirebbe necessariamente per andare ad avere delle ripercussioni brutali sulla mia produttività, sulla produttività di quello che cerchiamo di fare con l'azienda. Dire che l'AI moderna produca slop è non avere idea della cosa di cui si parla, perché oggi come oggi un AI guidata solo a livello progettuale esterno, ma non che gli devi dire di preciso, le idee proprio architetturali di base e di prodotto, produce per lo più codice migliore delle prime iterazioni umane di programmatori capaci, perché che nello stesso tempo lei ha reiterato molte più volte, è andata molto più in profondità, [...] Ma se tu invece segui il progetto e tutto il codice è scritto da un AI di frontiera come Astra, quello che viene fuori non è slop, è un programma spesso di ottimo livello, a volte di livello eccezionale e così via.*

Fonte: https://www.youtube.com/watch?v=sOt0WpQG5eU

Effetto «human bottleneck»







A hit on personal focus and productivity

Software engineer Dima Zaytsev (currently at Linear, formerly my colleague at Uber) told me something that felt very relatable on AI and productivity:

*"There's lots of context switching: **there's always another agent that waits on your response**.*

Also, it's just more work. When AI was ramping up, I felt more productive than my peers: I did in an hour what others needed a day for. Now, the expectation has become that everyone works on parallel things. And so, I end up feeling I have more work compared to pre-AI."

https://open.substack.com/pub/pragmaticengineer/p/the-state-of-the-tech-industry-in



Why Software Engineering Is Indispensable in the Age of Coding Agents

ALFONSO FUGGETTA, Politecnico di Milano, Italy

Can AI make Software Engineering (SE) — the discipline — obsolete? And can it make software engineers — the professionals — redundant? This paper argues that the rise of capable AI coding agents makes SE and software engineers essential, not obsolete: the missing foundation without which AI-assisted development produces misleadingly plausible, unverifiable, and ultimately untrustworthy software. Three structural properties of large language models (probabilistic generation, agnosticism, and semantic statelessness) create a structural vacuum that no amount of training can eliminate. Filling it requires four knowledge levers: methodological knowledge, domain knowledge, design choices, and process choices. All four must be reified as persistent artifacts, and each requires the software engineer as methodologist, mediator, and custodian.

CCS Concepts: • **Software and its engineering** → **Software creation and management**; • **Computing methodologies** → *Artificial intelligence.*

Additional Key Words and Phrases: software engineering, AI coding agents, large language models, software process governance, specification, vibe coding

https://www.abassavoce.it/p/un-mio-articolo-su-communications

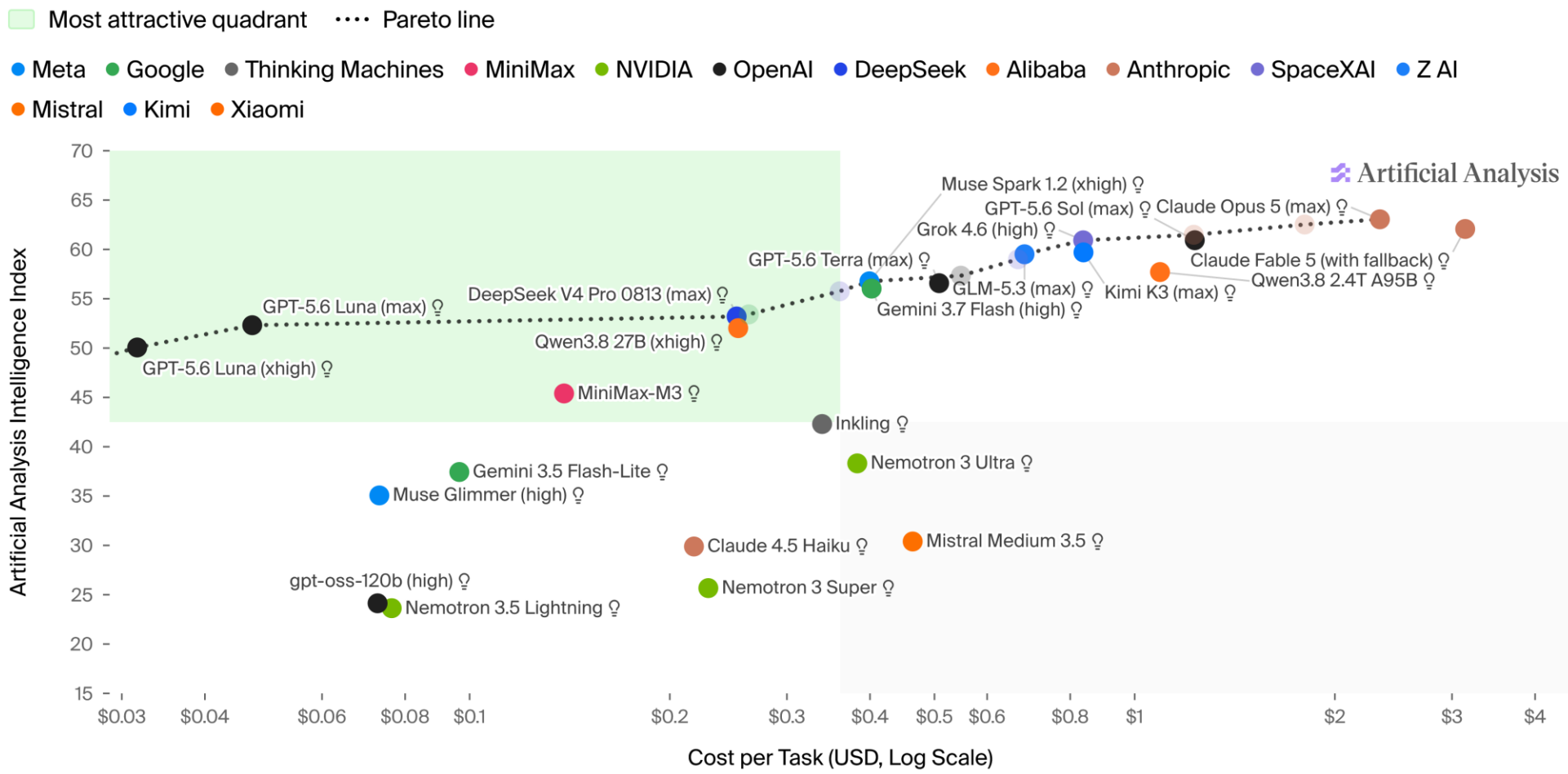
Mie personalissime conclusioni

Approccio Spec-Driven Development (SDD)

- Funziona! La mia sensazione è che, **per software di questo tipo**, siamo arrivanti in sostanza al punto in cui ci si può «concentrare sulle idee e non sulla scrittura del codice» (cit. Salvatore Sanfilippo aka antirez; esattamente come nel suo caso con DwarfStar4, al momento per Retriva non è stato scritto codice manualmente)
- Nessuna contraddizione con le buone pratiche consolidate da decenni nell'ingegneria del software (es. "Legge di Boehm"): oggi ha ancora più senso allocare molto del proprio budget temporale nella fase di concepimento e design
- Prestare molta attenzione alla selezione dei modelli. Intorno ai modelli «di frontiera» c'è moltissimo hype per ovvi motivi commerciali. A mio parere, il loro elevato costo è giustificato solo per task particolarmente complessi per cui effettivamente sia necessaria la performance più elevata disponibile nel mercato. Rule of thumb: in prima battuta, mantenersi in prossimità della frontiera di Pareto prestazioni/costo.

Intelligence Index vs. Cost per Intelligence Index Task

Artificial Analysis Intelligence Index · Weighted average cost (USD) per Artificial Analysis Intelligence Index task



Reasoning models are indicated by a lightbulb icon

Mie personalissime conclusioni

Molteplicità dei modelli, protezione dei dati

- Modelli per uso «funzionale»: non esiste un modello buono per tutte le stagioni; è opportuno stabilire un mix in funzione del caso d'uso e degli utilizzi all'interno dell'applicazione. Privilegiare implementazioni flessibili, ovvero che consentano facilmente di passare da un LLM ad un altro.
- Protezione dei dati: anche in questo caso non siamo in una situazione *one size fits all*, ovvero il tipo di deployment va scelto in base al caso d'uso. Di nuovo, consiglio caldamente di adottare un'architettura flessibile in modo da potersi adattare virtualmente a tutti gli scenari.



Per maggiori info

am.dev.75@gmail.com

https://www.linkedin.com/in/marsonandrea/

https://youtube.com/playlist?list=PLIXdajtWCbM6ChFh6P8dML1I9uAQhjOZ9

// Valuta questa sessione, grazie!



Andrea Marson

Agente conversazionale per uso aziendale/professionale basato su RAG e infrastruttura confidential computing



CONTATTI

hello@cloudconferenceitalia.it
www.cloudconferenceitalia.it/
www.linkedin.com/company/cloud-conference-italia

Grazie!