Virtual Assistant & AI Agent



Camera di Commercio
Venezia Giulia
Trieste Gorizia

punto impresa digitale

isonlab

Agenda

- Introduzione
- Tecniche
- Sistemi
- Training
- Ottimizzazione
- Funzioni attive
- Interazioni
- Pro/Cons
- Q&A



Introduzione



Intelligenza Artificiale:

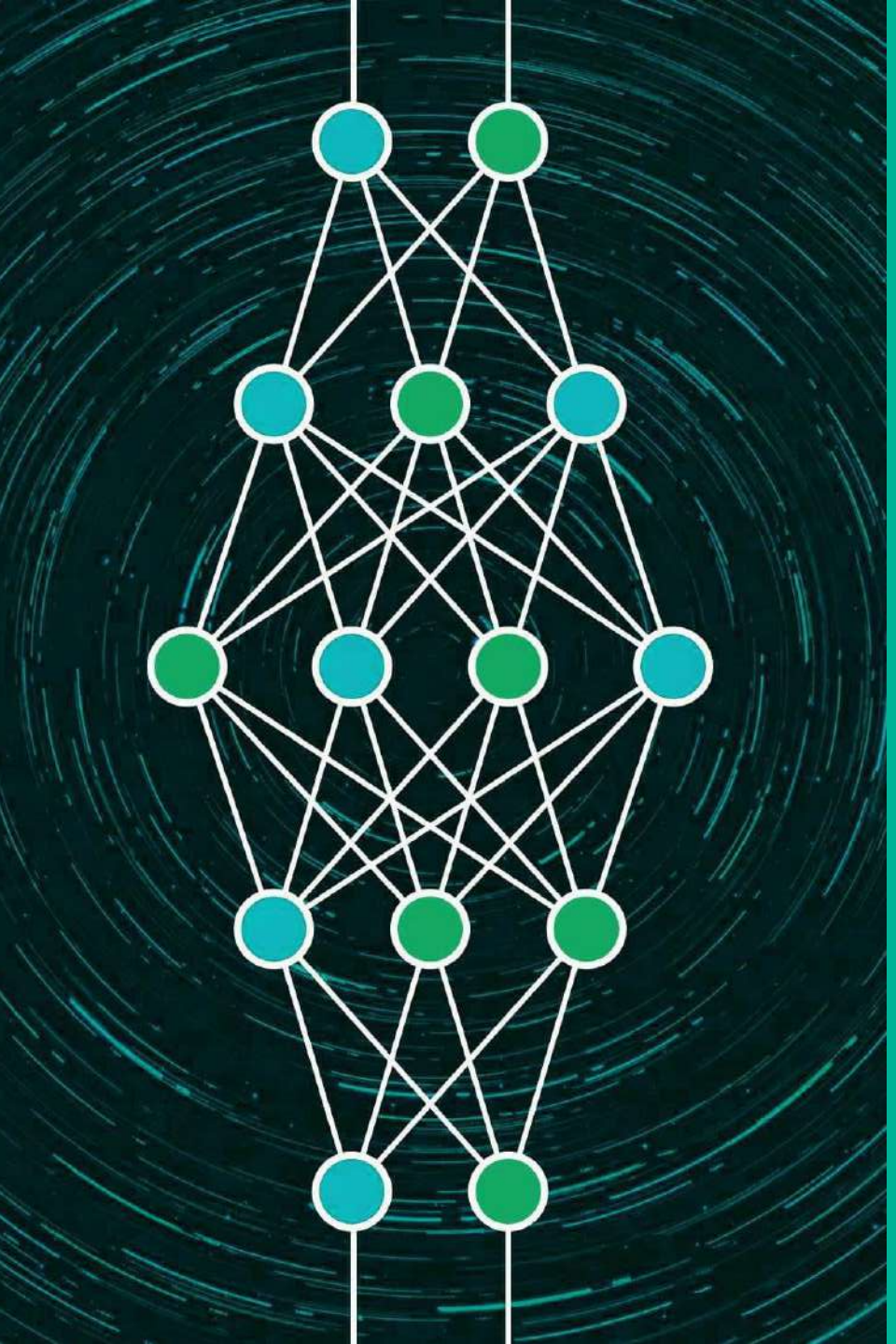
qualsiasi algoritmo che permetta ad un computer di simulare la razionalità umana

Machine Learning:

metodi statistici che permettono il miglioramento della performance attraverso l'esperienza

Deep Learning:

reti neurali a ragionamento gerarchico multilayer, apprendimento non supervisionato e astrazione



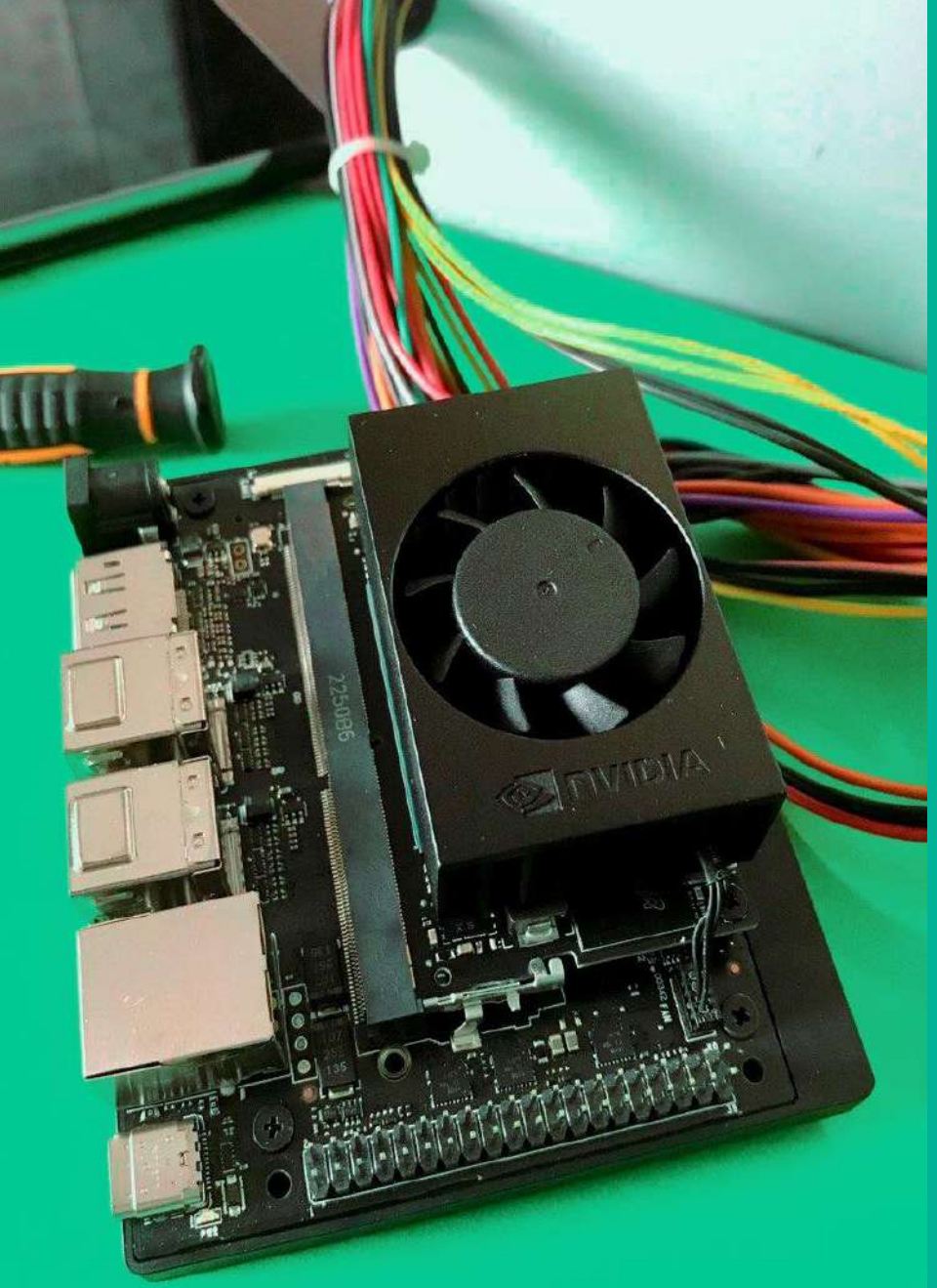
Reti Neurali

Le reti neurali sono delle *black box*.
Non sappiamo perché raggiungano queste performance, ma sappiamo come lo fanno

I cosiddetti *dense layers*, che sono gli strati più comuni in una rete neurale, creano le interconnessioni tra i vari strati della rete

Ogni neurone è connesso a tutti quelli dello strato successivo, il suo valore di output diventa quindi input per i prossimi neuroni.

Ogni connessione tra neuroni possiede un peso (*weight*) che è uno dei fattori che viene modificato durante l'addestramento.



HARDWARE

In quanto a processori, oggi le GPU, ad alto parallelismo, hanno soppiantato le CPU in funzioni computazionali.

NVIDIA Jetson Orin Nano

(275 trilioni di operazioni al secondo)

- elabora input multi-modali (voce, testo, immagini)
- tempo reale e multiutente
- altamente scalabile



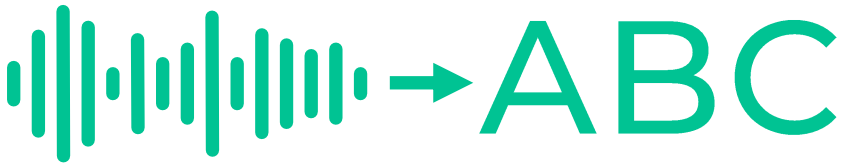
Cosa è l’IA Conversazionale?

L'intelligenza artificiale conversazionale è una tecnologia che permette alle macchine di interagire con gli esseri umani attraverso il linguaggio naturale, simulando una conversazione simile a quella umana.

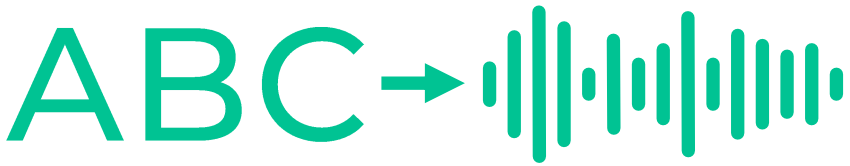
Si basa su diversi approcci tecnologici che le permettono di comprendere, elaborare e generare risposte in modo pertinente e in tempo reale.

Tecniche

STT/TTS



Speech-to-Text (STT): consente alle macchine di interpretare il linguaggio umano parlato, e quindi i comandi vocali.



Text-to-Speech (TTS): abilita le macchine a "parlare" in linguaggio naturale, offrendo esperienze di interazione fluide e realistiche.

Permettono agli utenti di interagire vocalmente con i vari modelli, rendendo le tecnologie di linguaggio pienamente accessibili a chiunque.

NLP

Definizione:
Il ***Natural Language Processing*** (NLP) è il ramo dell'intelligenza artificiale che si occupa dell'interazione tra computer e linguaggio umano. Consiste in metodi e tecniche per permettere alle macchine di comprendere, interpretare e generare testo in modo significativo.

Obiettivi:
Comprensione testuale: analizzare e interpretare il significato di frasi e testi
Categorizzazione: identificare e classificare in grandi volumi di testo scritto
Traduzione automatica: tradurre testi da una lingua all'altra
Riconoscimento: identificare nel testo entità (nomi, luoghi, etc)
Sentiment Analysis: comprendere le emozioni che traspaiono dai testi

LLM

Definizione:
I ***Large Language Model*** (LLM) sono un forma avanzata di modelli del dominio NLP, addestrati su enormi quantità di testo per generare risposte contestuali, completare frasi, e comprendere input complessi.

Caratteristiche:
Dimensioni: contengono miliardi di parametri, per una comprensione avanzata
Generalizzazione: capacità estese di traduzione, riassunto, ricerca, etc

Implicazioni:
Sono strumenti potenti declinabili ad ambiti complessi come automazione, assistenza, produzione di contenuti e sviluppo di tecnologie di conversazione.



NLP

Un bibliotecario efficiente, competente e veloce.

"Trova un libro sulla storia di Roma": sa trovare il libro giusto, basandosi sul titolo o sugli argomenti principali che lo riguardano.

"Quali sono i capitoli più importanti?": sa rispondere correttamente, riassumere e consigliarti libri simili.

Risposte dirette a domande mirate. Sa dove trovare le informazioni che servono in modo rapido.

LLM



Un bibliotecario esperto e chiacchierone, ha letto tantissimi libri, riviste e articoli di ogni tipo.

"Raccontami la storia di Roma": risponde con una spiegazione dettagliata e completa. Sa molto di storia ma anche di tanti argomenti relati, come filosofia, arte e la cultura dell'epoca.

"Come sarebbe oggi la società moderna se l'Impero Romano non fosse caduto?": risponde basandosi su tutto ciò che ha imparato in un contesto ampio.

Risponde, discute, approfondisce, racconta dettagli e crea connessioni speculative tra argomenti diversi.

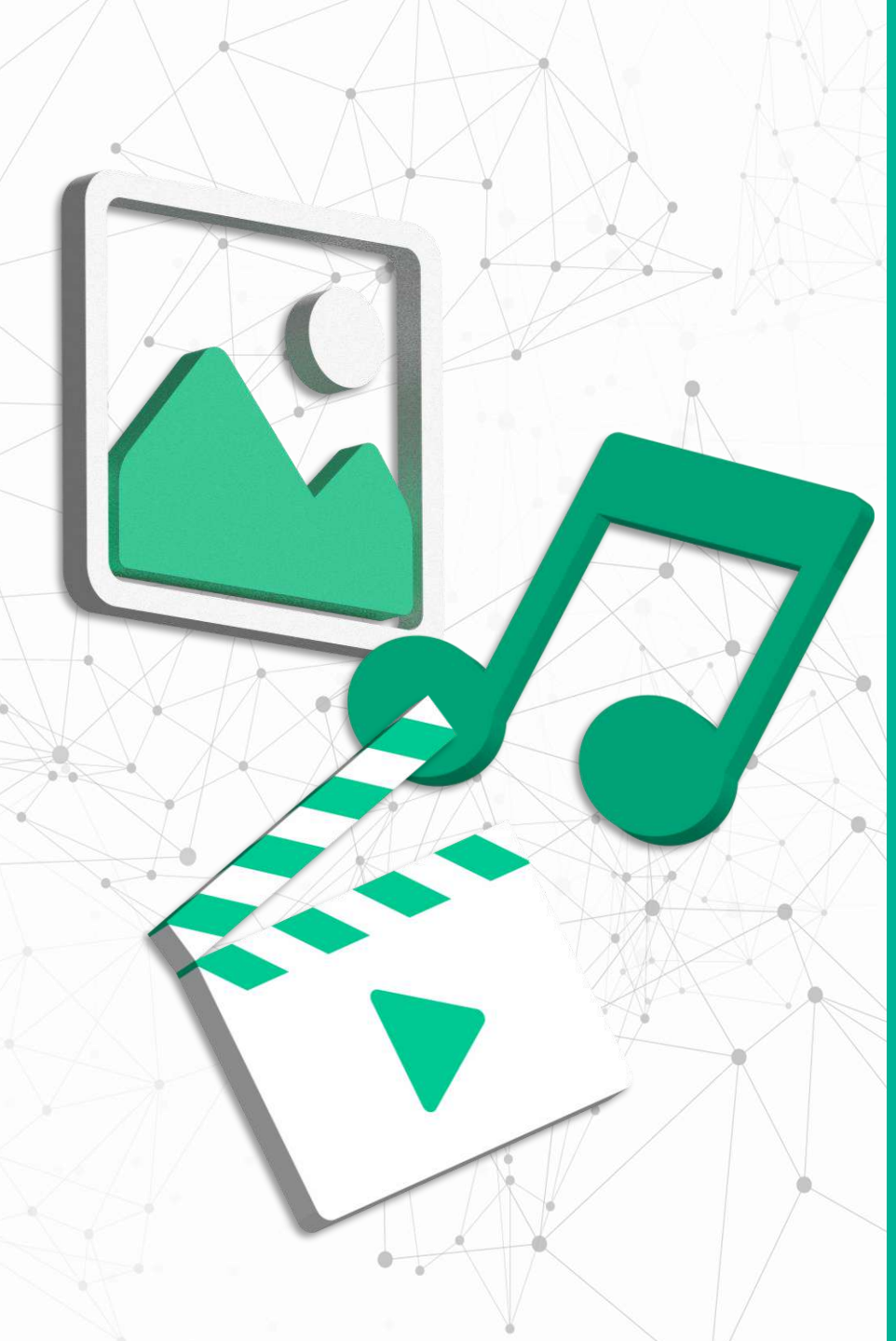
Modelli generalizzati

Sono in grado di svolgere vari compiti linguistici senza essere specificamente addestrato per ciascuno. Sono modelli multiuso, flessibili e altamente scalabili.

Maggiore versatilità: svolgono compiti diversi con un unico sistema, adattandosi a traduzione, riassunto, risposta a domande, e generazione di contenuti senza bisogno di addestramento specifico per ogni compito.

Maggiore Adattamento: si adattano a settori specifici tramite tecniche di fine-tuning, riducendo tempi e costi di personalizzazione.

Riduzione dei bias: Hanno meccanismi per riconoscere e mitigare i bias attraverso una maggiore varietà di dati di addestramento e controlli continui.



I modelli generalizzati sono un concetto estremamente ampio e pervasivo.

Possono essere LLM, ma il termine può anche includere modelli multimodali (capaci di lavorare con immagini, testo, suoni) che non si limitano esclusivamente al linguaggio naturale.

Sistemi

Chatbot

Progettati per simulare l'interazione di una conversazione con un essere umano attraverso chat testuali o vocali per rispondere a task standardizzate, eseguire azioni semplici e ripetitive. Utili per gestire FAQ, assistenza base o raccolta dati.

Caratteristiche:

Semplicità: Facili da integrare, risposte preimpostate per conversazioni lineari

Task basilari: Ideali per domande standard ricorrenti (info su orari, costi, etc)

Supporto 24/7: Disponibilità ininterrotta per i clienti.

Capacità limitate: non sono in grado di apprendere dall'esperienza.

Soluzione facilmente integrabile, semplice e a basso costo, funzionale alle realtà che gestiscono volumi elevati di interazioni con richieste prevedibili e ripetitive, che non necessitando di una profonda comprensione del contesto

Chat History 15

Search Chat Group

Audio Generate Just now

Design thinking is a problem-solving approach .....

2:40

E-Commerce Website Code 2 Min ago

Design thinking is a problem-solving approach.....





Design Thinking 1 Hr ago

Design thinking is a problem-solving approach .....

E-Commerce Website Code Just now

Design thinking is a problem-solving approach.....






Add New Group

Photo Editing Yesterday

Design Thinking



2 Mini ago

Tell me about Design Thinking.

Save Question | Copy Question | Edit

3/3

2 Mini ago

Design thinking is a problem-solving approach that emphasizes empathy, creativity, and collaboration. It involves understanding the needs and perspectives of users, generating innovative ideas, and rapidly prototyping and testing solutions.

Copy Text | Regenerated Response | Copy Link | Share



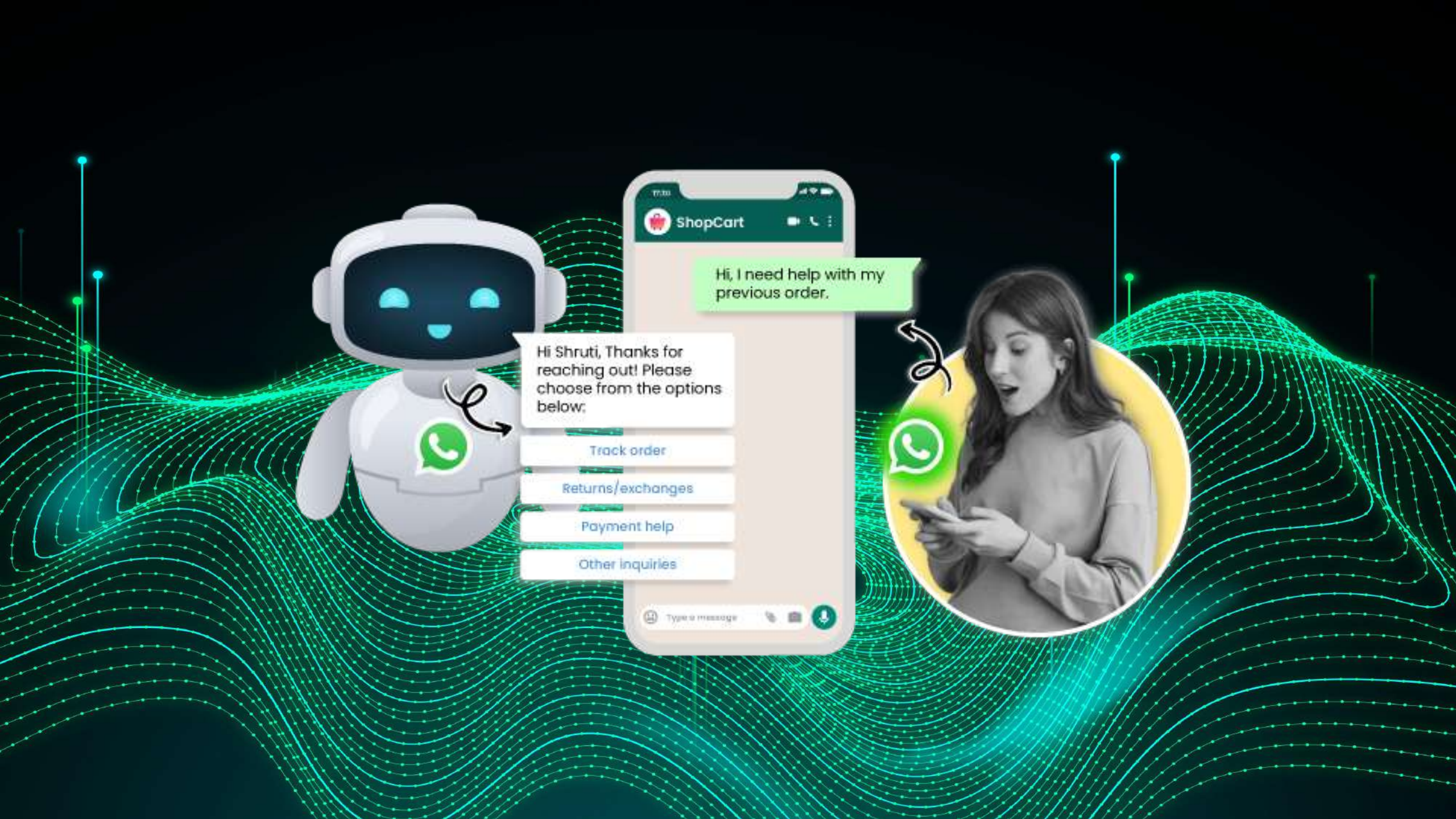
2 Mini ago

What are the key principles of design?

Save Question | Copy Question | Edit

Prompt | Upload File

Ask Me anything ....


ShopCart
Hi, I need help with my previous order.
Hi Shruti, Thanks for reaching out! Please choose from the options below:
Track order
Returns/exchanges
Payment help
Other inquiries

Assistenti Virtuali

Usano NLP e Machine Learning per comprendere il linguaggio umano, offrendo risposte fluide e personalizzate. Non seguono uno script predefinito, possono comprendere il contesto e apprendere dai dati per migliorare nel tempo.

Caratteristiche:

Interazione contestuale: si adattano in tempo reale

Personalizzazione: accedono a dati storici per offrire risposte su misura

Automazione avanzata: gestiscono attività complesse

Esperienza: l'interazione evoluta migliora la soddisfazione degli utenti

Eg. *Siri* e *Alexa*. Funzionali a combinare automazione con personalizzazione in interazioni contestuali. Non solo rispondono a domande dirette, ma eseguono azioni complesse come controllare dispositivi o inviare messaggi.




2:56
PM
TA
NAVIGATION
Mount Street
London
PHONE
Sara's Phone
MEDIA
Young Before
Your Time
Corvidee
AUTO
SYNC
AUTO
MAX A/C
A/C
MAX
Push
Pull
19.5
Push
Pull
20

AI Agent

La forma di evoluzione più avanzata di queste tecnologie, sono progettati per operare reattivamente, ricordare il pregresso e prendere decisioni complesse in modo pienamente autonomo all'interno di ambienti altamente dinamici.

Caratteristiche:

Autonomia: non necessitano di input, possono prendere decisioni e agire (ML)

Apprendimento continuo: reinforcement learning per migliorare le prestazioni

Interazione multi-dominio: operano su diversi fronti contemporaneamente

Adattività in tempo reale: risposte rapide e reazione al cambiamento

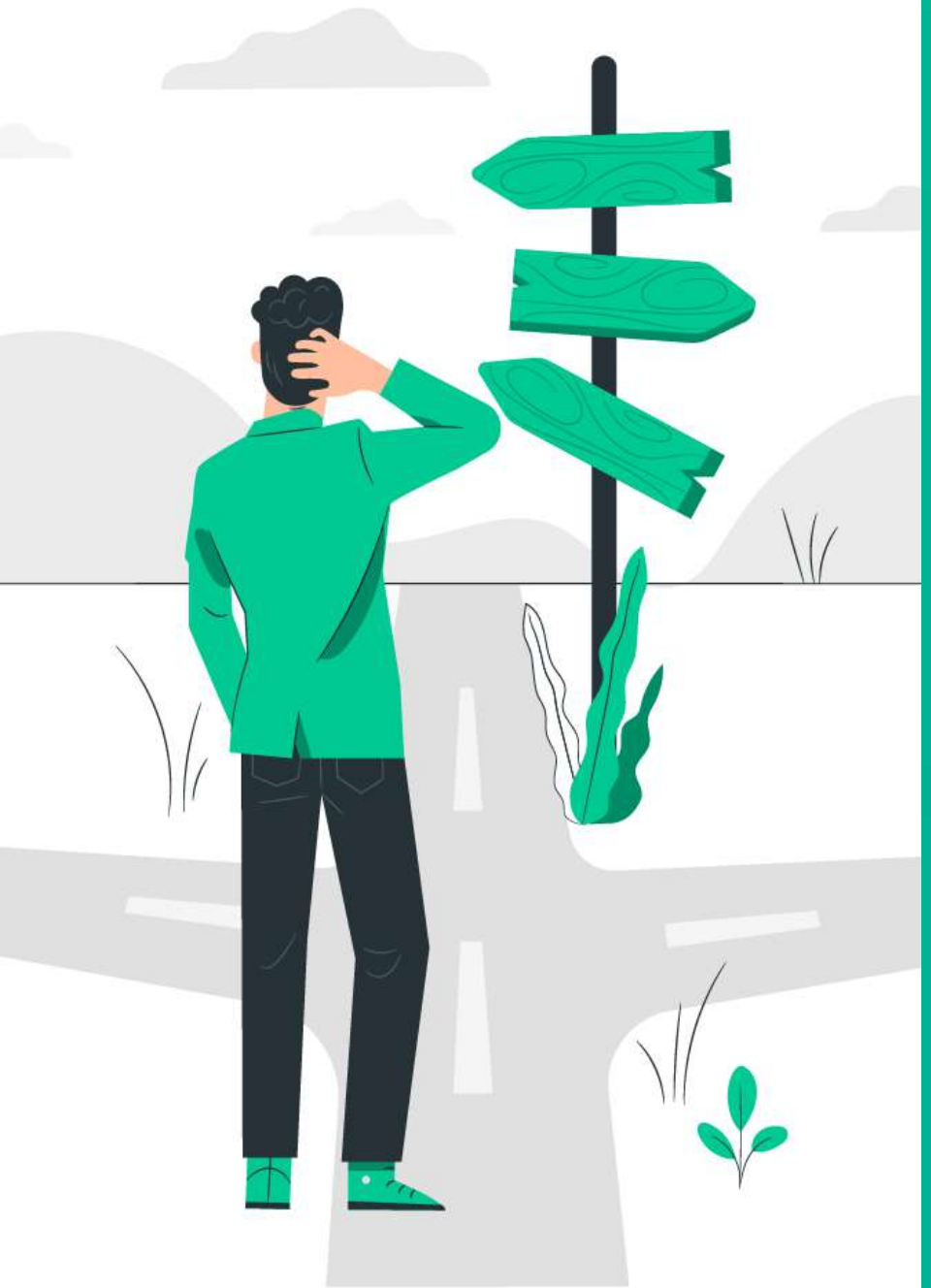
Ideali per ambienti dinamici dove sono necessari processi decisionali complessi e reattivi che richiedono livelli di automazione intelligente (eg. risorse umane, manifattura, settore finanziario, logistica).






oculus

	CHATBOT	ASSISTANT	AI AGENT
obiettivo	Simulare conversazioni umane	Assistenza con compiti specifici	Sistema intelligente e autonomo
interattività	Basato su dialoghi strutturati	Interazioni guidate dall'utente	Agisce e interagisce in autonomia
complessità	bassa	media	elevata
proattività	reattivo	reattivo	proattivo
personalizzazione	minima	limitata	elevata
apprendimento	basso: script o modelli preaddestrati	limitato: apprende da input espliciti	elevato: tecniche di ML per automigliorarsi



Quale tecnologia scegliere:

Non c’è una risposta corretta in senso assoluto.
dipende direttamente dalle esigenze specifiche.

Condividono alcuni elementi comuni, mentre le differenze risiedono nella complessità e capacità di apprendimento/autonomia.

Tutto parte dalla consapevolezza.
Identificare la tecnologia giusta non solo migliora l’efficienza operativa, ma fornisce un vantaggio competitivo, migliora l’esperienza dei clienti e riduce i costi operativi nel lungo termine.

Training

Training base

Input iniziale:
Si raccolgono grandi quantità di conversazioni, testi o dati pertinenti all'ambito di utilizzo (eg. trascrizioni di chat, documentazione tecnica, FAQ).

Pulizia e Preprocessing:
Si rimuovono dati rumorosi, formattando le informazioni per renderle uniformi. (eg. eliminazione di duplicati o rimozione di informazioni sensibili)

Adattamento:
Si parte da un modello pre-addestrato (come GPT o BERT) che ha già imparato le basi del linguaggio naturale.
Il modello viene quindi adattato in modo circoscritto a dati specifici

Architettura

Si definisce l’architettura conversazionale.

Intents: Identificare le intenzioni degli utenti
- eg. *"Controllare gli ordini", "Risolvere un problema tecnico".*

Entities: Riconoscere informazioni chiave nella frase
- eg. *"controllo ordini in entrata" : "ordini i entrata" è un’entità*

Dialog Flow: *Progettare il percorso di conversazione*
- eg. " vuoi sapere altro sui tuoi ordini?"

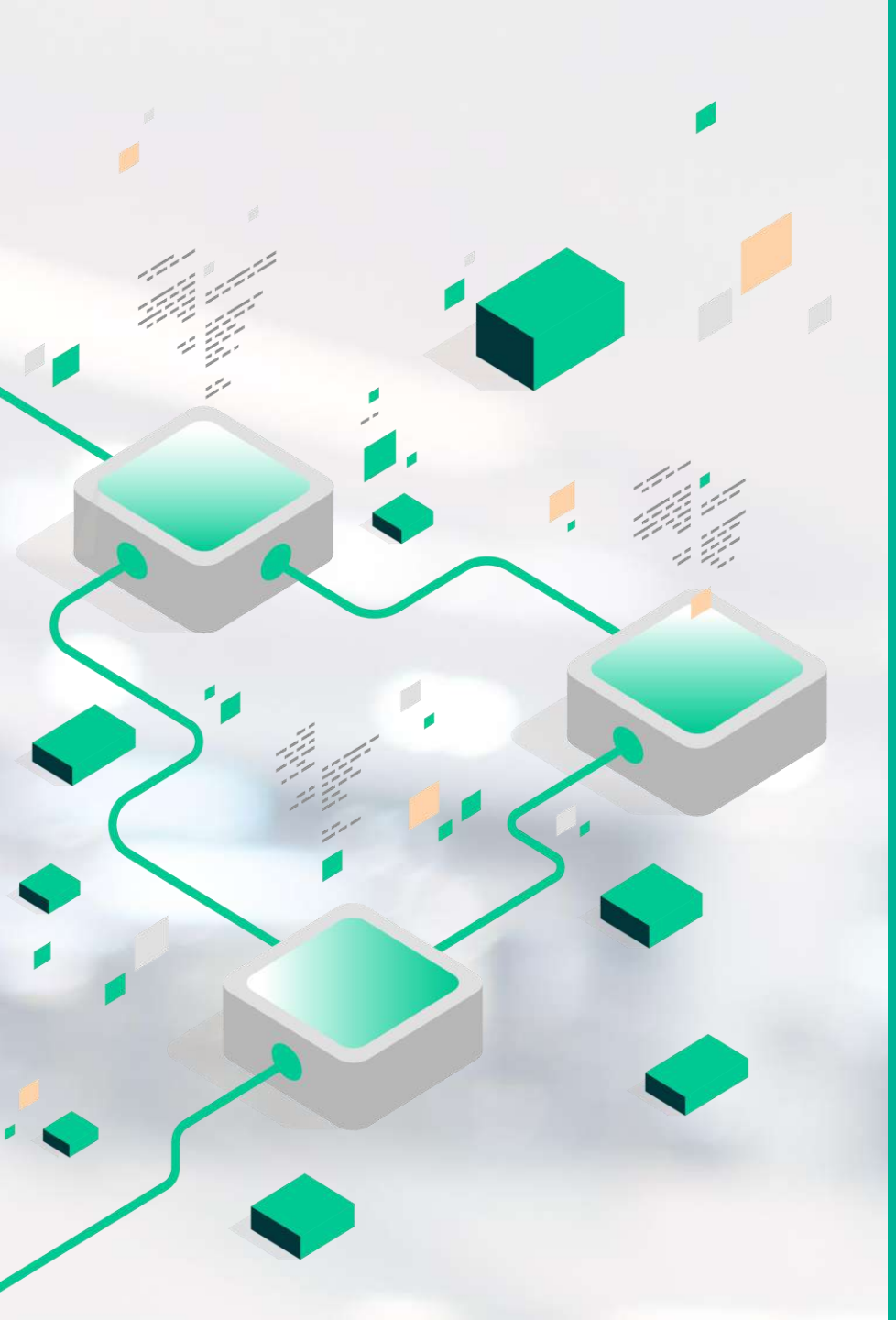
Test

Questo ciclo garantisce che il modello evolva per offrire risposte sempre più precise e utili, adattandosi ai bisogni dell'utente.

Simulazioni: test su conversazioni simulate
- eg. Un tester chiede "*Come faccio a cambiare la mia password?*", e si verifica la risposta del modello.

Feedback umano: utenti o annotatori valutano la qualità delle risposte, fornendo miglioramenti.

Aggiustamenti: ottimizzazione e modifiche (modello/dati) per ridurre errori
- eg. Se molti utenti chiedono una funzione non prevista, i dati vengono aggiunti al training.



REINFOCEMENT LEARNING

Tecnica per realizzare agenti autonomi in grado di scegliere azioni da compiere per raggiungere determinati obiettivi interagendo con l'ambiente in cui sono immersi.

La qualità di un'azione è data da un valore numerico di "ricompensa« che ha lo scopo di incoraggiare comportamenti corretti dell'agente

Si occupa di problemi di decisioni sequenziali, in cui l'azione da compiere dipende dallo stato attuale del sistema e ne determina quello futuro.

Ottimizzazione



Le capacità e l'efficacia di un modello IA sono fortemente determinate dal tipo addestramento.

A partire dallo specifico LLM, esaminiamo due diverse possibilità di migliorare le prestazioni di una intelligenza artificiale generativa.

- **Fine Tuning**
- **RAG (Retrieval Augmented Generation)**

FINE TUNING

Il processo si riferisce all'adattamento di un modello IA già pre-addestrato a un compito specifico o a un set di dati particolare (*LoRa* - Low Ranking).

In pratica, il modello, che è stato addestrato su un vasto e generico dataset, viene ulteriormente addestrato (o "affinato") su un dataset più piccolo e specifico.

Applicabile a modelli di ogni dimensione, questo consente una maggiore specializzazione su certi compiti o tipo di dati.

Viene spesso utilizzato per migliorare le prestazioni in specifici ambiti o per adattare il modello a linguaggi, stili o formati particolari.

Efficienza in argomenti specializzati: In contesti altamente specializzati, il corpus di conoscenze rilevanti è spesso più ristretto e meno variabile. Un modello più piccolo, fine-tuned su un dataset specifico, può imparare efficacemente senza la necessità di gestire l'ampia gamma di informazioni che un modello più grande deve affrontare.

Velocità e risorse: Modelli più piccoli richiedono generalmente meno risorse computazionali e di memoria. Questo li rende più veloci da addestrare e da eseguire, un vantaggio significativo in applicazioni pratiche che richiedono risposte rapide.

Minore rischio di Overfitting: In contesti molto specifici, un modello più grande può a volte soffrire di overfitting, ovvero diventare troppo adattato al dataset di training e quindi meno in grado di generalizzare o di gestire variazioni non presenti nel training. Un modello più piccolo ma focalizzato riduce invece questo rischio.

Precisione nel contesto specifico: Un modello fine-tuned su un dataset specializzato può diventare molto preciso nel contesto per cui è stato addestrato. Se il dataset è rappresentativo dell'uso previsto del modello, quest'ultimo fornisce risultati di alta qualità per quella specifica applicazione.

RAG

Approccio ai modelli generativi, combina la generazione di risposte con un meccanismo di ricerca di informazioni.

In pratica, quando il modello deve generare una risposta, prima esegue una ricerca nelle sue fonti di dati (come documenti, pagine web, ecc.) per trovare informazioni pertinenti. Usa queste informazioni recuperate per informare o arricchire la risposta generata.
Particolarmente utile per aumentare l'accuratezza e la pertinenza in contesti dove sono necessarie informazioni specifiche o aggiornate.

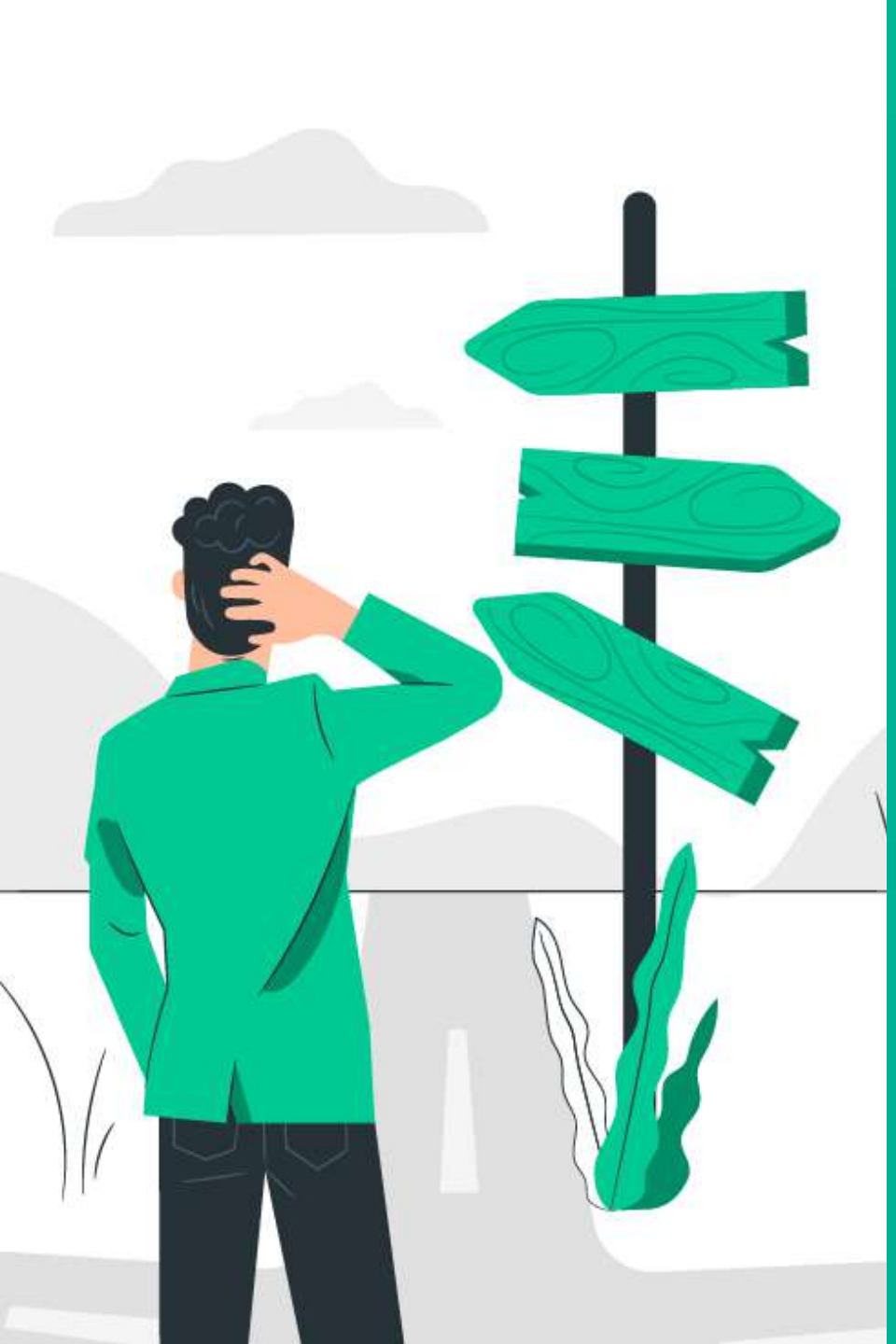
L'approccio RAG è più complesso e adatto a grandi modelli. Non risponde basandosi solo sul training, ma integra la capacità di ricercare e utilizzare informazioni esterne. Adatto a domande molto specifiche e diversificate, che richiedono risposte basate su informazioni recenti e mutevoli.

Informazioni aggiornate: molti settori sono soggetti a cambiamenti frequenti come nuove scoperte o prodotti, variazioni dei mercati e cambiamenti normativi. Un sistema RAG può recuperare informazioni da fonti esterne aggiornate per fornire risposte che riflettano le ultime novità e i dati più recenti.

Ampie competenze: in un settore ampio e complesso, gli utenti possono avere domande che spaziano da richieste tecniche specifiche a richieste di consigli generali. Grazie alla vasta gamma di fonti, si gestisce questa diversità in modo più efficace rispetto a un modello basato solo sul fine tuning.

Minore necessità di aggiornamenti: attingendo automaticamente alle ultime informazioni disponibili online, si riduce la necessità di addestramenti frequenti per garantire una conoscenza aggiornata.

Gestione di domande non prevedibili: in campi a rapida evoluzione, spesso emergono nuove questioni e argomenti che non erano state precedentemente considerate. Il RAG permette di gestirle in modo efficace.



Quale approccio per le PMI:

Non c’è una risposta corretta in senso assoluto.
L’approccio tecnico deve essere sensibile al caso specifico per offrire:

- Velocità e fluenza della conversazione
- Profondità e specificità delle informazioni
- Coerenza con *tone-of-voice*
- Ottimizzazione energetica computazionale
- Gestibilità e sicurezza

Tutto parte dalla consapevolezza.

Creare base dati proprietari circoscritta ad ambito specifico e integrarla con modello esistente per una soluzione snella, facilmente gestibile ed efficiente con maggiore controllo sulla privacy e sulla sicurezza dei dati.

Funzioni Attive



Function Call

Il termine fa riferimento a una capacità avanzata del modello di linguaggio, che gli consente di invocare funzioni specifiche per rispondere a richieste dell'utente o per eseguire azioni particolari. Queste "chiamate di funzione" possono essere definite come delle API interne o esterne che il modello richiama per ottenere informazioni o per eseguire operazioni che non è in grado di compiere autonomamente.

Esempi Function Call

Attività di scraping: massima autonomia nella ricerca dati all'interno di database, documenti, siti web, etc. per trasformarli in un formato strutturato per analisi e modellazione, migliorando la qualità e l'affidabilità delle informazioni fornite.
Esempio: *"Chi ha vinto le elezioni USA?"*

Interazione con sistemi esterni: se l'utente desidera eseguire un'azione (eg. modificare un ordine, aggiornare dati o prenotare un servizio), l'IA può utilizzare una Function Call per inviare la richiesta al sistema dedicato.
Esempio: *"Vorrei aggiornare l'indirizzo di consegna"*

Accesso a informazioni specifiche: se l'IA deve accedere a dati specifici (come l'ultimo acquisto o i dati di registrazione), può fare una Function Call al database dell'azienda per recuperare le informazioni in modo sicuro.
Esempio: *"Quando è programmata la prossima consegna?"*



Memory & Recall

In ambito di intelligenza artificiale e, nello specifico, nell'elaborazione del linguaggio naturale, le funzioni di Memory (memoria) e Recall (richiamo o recupero) si riferiscono a capacità specifiche:

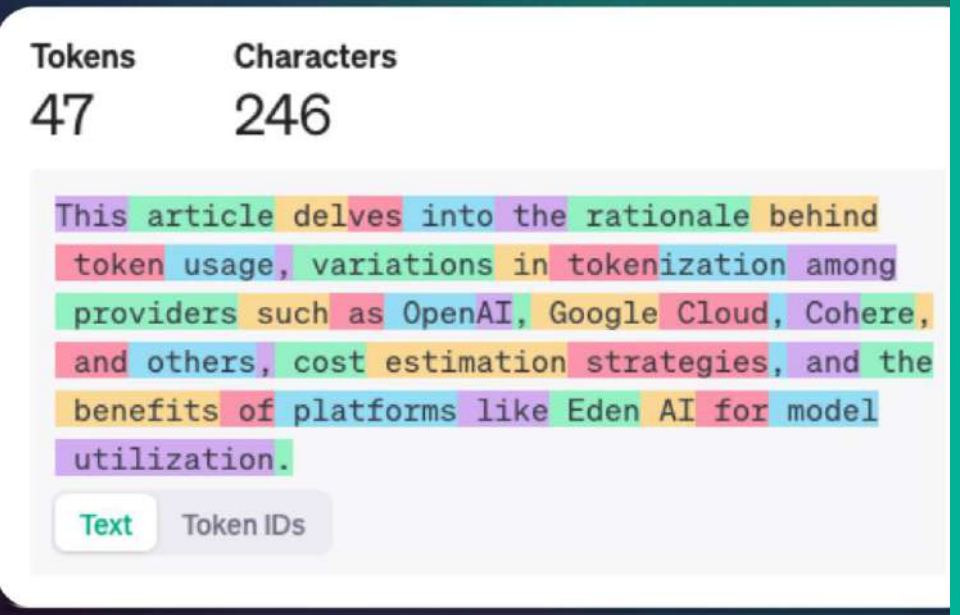
- memorizzare concetti spiegati in precedenza e utilizzarli per costruire argomenti successivi
- richiamarle le informazioni apprese in modo efficace quando necessario

Esempi Memory & Recall

***Conversazioni contestuali**: l'IA può ricordare messaggi precedenti o argomenti già discussi. Un app di tutoraggio può ricordare quali concetti l'utente ha già appreso e quali necessita di approfondire.*

***Assistenza personalizzata:** In una piattaforma di assistenza, la memoria consente all'IA di conservare dettagli sull'utente o sui problemi precedenti per fornire risposte personalizzate, senza richiedere di ripetere nuovamente gli input.*

***Raccomandazioni:** Un sistema può usare il recall per proporre contenuti simili a quelli che l'utente ha richiesto o dimostrato di apprezzare in precedenza, migliorando la personalizzazione e l'esperienza complessiva.*



Tokenizzazione

Definisce il modo in cui il testo viene suddiviso in unità comprensibili dette "token«, che il modello può elaborare e memorizzare.

Permette di gestire una grande quantità di testo e ottimizza la capacità di memoria e di recall.

Tokenizzare termini come “AI”, “IA” e “Intelligenza Artificiale” in una sola rappresentazione non duplica concetti e migliora la coerenza nelle risposte.

Interazioni

Interazione Utente

Comunicazione umanizzata:

- Capacità di comprendere e rispondere in linguaggio naturale
- Flusso bidirezionale

Integrazione di documenti e immagini:

- Gli utenti possono inviare documenti e immagini
- Analisi e gestione documentale per risposte più accurate

Compatibilità con sistemi XR:

- Possibilità di utilizzo in contesti di realtà estesa (XR)
- Supporto alle applicazioni per maggiore usabilità

Interazione Multi Dominio

- ***Gestione delle attività:***
  Esempi di automazione delle attività ripetitive e dei flussi di lavoro

- ***Gestione risorse umane:***
  Uso per la selezione, l'onboarding, il training e la gestione HR

- ***Routing delle richieste e assistenza operativa:***
  Possono ottimizzare il flusso delle richieste all'interno di un'organizzazione

- ***Gestione asset:***
  Monitoraggio, tracciamento e gestione degli asset tramite AI Agents

Pro/Cons

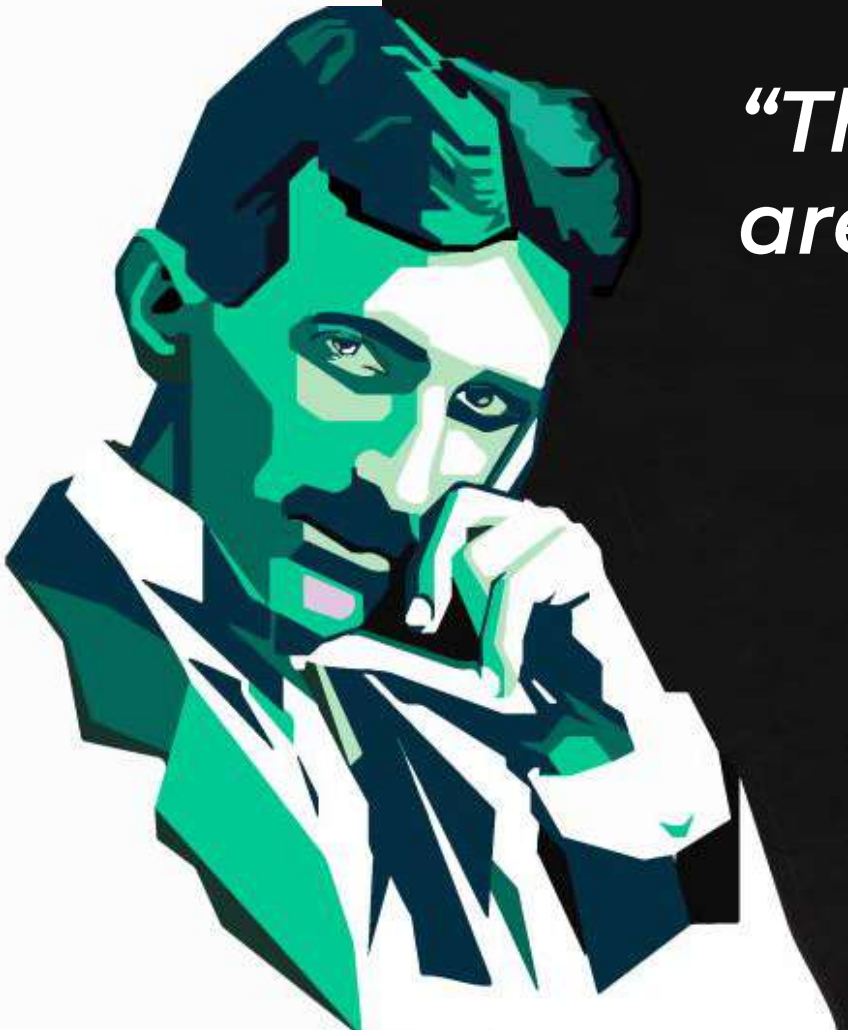
Vantaggi e Sfide



- aumento della produttività
- riduzione dei tempi di risposta
- migliore qualità del servizio

- coerenza tecnologica
- competenza nell’integrazione
- assicurare eticità e privacy

Q&A



“The wonders of yesterday are today common occurrences”

Nikola Tesla

isonlab

dott. Francesco Passador

Responsabile Marketing
francesco.passador@isonlab.it