IA PER LA COMUNICAZIONE

Large Language Models (LMM)

Prof. ssa Romina Eramo
Dipartimento di Scienze della Comunicazione
Università degli Studi di Teramo
reramo@unite.it

Definizione

- Un Modello Linguistico di Grandi Dimensioni (LLM) è una rete neurale addestrata su enormi quantità di testo per elaborare, comprendere e generare linguaggio naturale.
- Gli LLM non "pensano" come un essere umano, ma hanno capacità che sembrano suggerire un'intelligenza simile.

Quanto è grande un LLM?

Due dimensioni da distinguere:

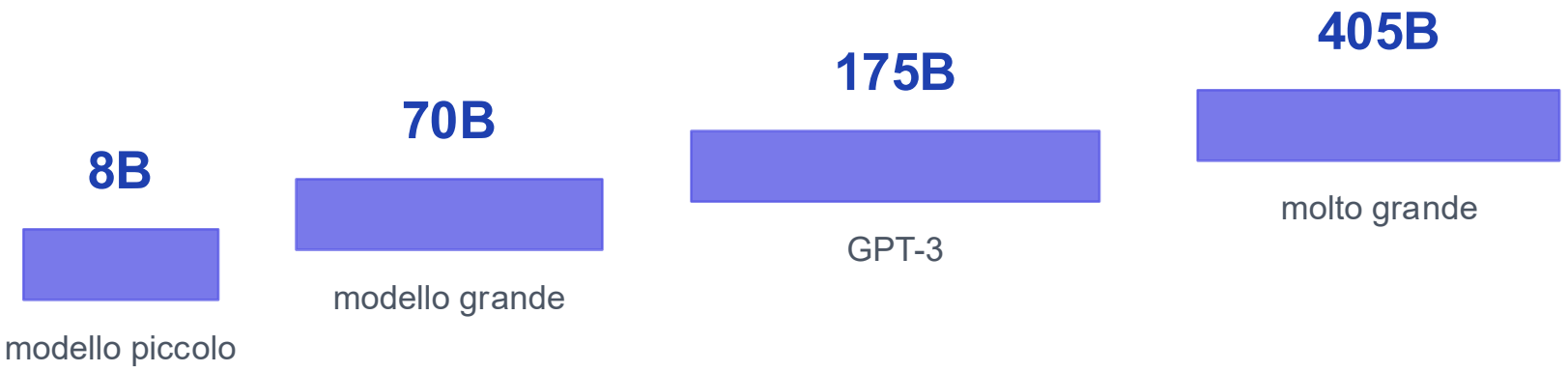
Parametri

Sono i “numeri” appresi dal modello durante l’addestramento.

Token di training

Sono i frammenti di testo usati per addestrare il modello.

Ordini di grandezza



Esempio: Meta Llama 3.1 dichiara versioni da 8B, 70B e 405B parametri e oltre 15 trilioni di token.

B = miliardi · T = trilioni

Dimensioni: cosa significano in pratica

1 parametro = un numero appreso

Non è una parola, una regola o un fatto: è una parte della configurazione interna del modello.

1 token ≈ pezzo di parola

I modelli non leggono direttamente "parole", ma sequenze di token: parole, parti di parole, segni, spazi.

Memoria dei soli pesi del modello

Stima con 2 byte per parametro (FP16/BF16):

8B	70B	405B
≈ 16 GB	≈ 140 GB	≈ 810 GB

solo per memorizzare i pesi; esecuzione e contesto richiedono ulteriore memoria

Messaggio chiave

Più grande non significa automaticamente migliore: contano anche qualità dei dati, architettura, post-training, reasoning e strumenti.

Dimensioni: esempio

- **GPT-6 Astra**
  - Parametri: **non dichiarati pubblicamente**
  - Token di training: **non dichiarati pubblicamente**
  - Context window: **1,05 milioni di token**
  - Output massimo: **128.000 token**

GPT-6 Astra → modello flagship, “most capable”, per reasoning complesso, coding, ricerca e computer use.
GPT-6.1 Sol → modello più recente della famiglia, orientato a un miglior equilibrio tra capacità e costo.
GPT-6 Luna → variante più veloce/economica per carichi elevati.

ANI, AGI e modelli general-purpose

- ANI (Artificial Narrow Intelligence): Sistemi progettati per compiti specifici.
- AGI (Artificial General Intelligence): ipotesi di sistemi con capacita cognitive generali paragonabili a quelle umane; non e realizzata.
- LLM moderni: sistemi general-purpose capaci di operare su testo, codice, immagini, documenti e strumenti.

- Le prestazioni su test di ragionamento non implicano automaticamente comprensione umana o AGI.

Domanda guida: che cosa misurano davvero questi test?

Un salto indietro: LLM nel 2023

Perche conservare esempi di Bard, GPT-3.5 e GPT-4?

Baseline storica

Gli output delle prossime slide sono una fotografia delle prestazioni osservabili a marzo 2023. Non rappresentano i modelli attuali.

Metodo didattico

Ripetere oggi gli stessi prompt consente di osservare miglioramenti, limiti persistenti e cambiamenti nello stile delle risposte.

Cosa osservare

Correttezza, qualita della spiegazione, gestione dell’incertezza, aderenza al prompt, tono e livello di dettaglio.

Capacità Emergenti degli LLM*: Scrittura creativa

Prompt: *"Scrivi una favola della buonanotte su un orsacchiotto, un tappeto magico e la luna."*

Bard: Una storia semplice e diretta, adatta a bambini molto piccoli.

"Un orsacchiotto trova un tappeto magico e vola sulla luna per incontrare alieni amichevoli."

GPT-3.5: Aggiunge dettagli, come un giardino lunare curato dagli alieni.

"Timmy, l'orsacchiotto, esplora crateri e incontra amici alieni."

GPT-4: Usa un linguaggio poetico e descrittivo.

"Cuddles vola verso la luna, danzando con stelle scintillanti e giocando a nascondino tra i crateri."

Quale versione cattura meglio l'immaginazione? Perché?

* Tutti gli output rappresentano le prestazioni del modello a fine marzo 2023. Bard nel 2024 è diventato Gemini.

Capacità Emergenti degli LLM: Ragionamento Logico

Esempio: Teoria della Mente

Scenario: Mary dice ad Amanda di amare John. Una settimana dopo, cambia idea e si innamora di Ralph, ma Amanda non lo comunica a Kathy.

Domanda: *"Chi pensa Kathy sia l'amore di Mary?"*

Bard: Assume che Amanda abbia parlato a Kathy del nuovo amore (sbagliato).

GPT-3.5: Analizza correttamente le informazioni disponibili, ma lascia aperta l'interpretazione.

GPT-4: Risponde con precisione: "Kathy pensa ancora che Mary ami John."

Come si comportano gli LLM rispetto al ragionamento umano?
Dove potrebbero fallire?

Capacità Emergenti degli LLM: Ragionamento Probabilistico

Il Problema Linda (Modificato)

Prompt: Abner si è occupato attivamente di questioni di giustizia sociale e ha partecipato a manifestazioni contro la polizia. Cosa è più probabile che Abner sia: un meccanico o un meccanico attivo nel promuovere riforma della polizia?

Risposta corretta: È più probabile che sia semplicemente un meccanico, poiché l'opzione "meccanico attivo" è una congiunzione (meno probabile).

Risultati:

GPT-4 risponde correttamente e spiega la regola della congiunzione.

GPT-3.5 e **Bard** offrono risposte corrette ma motivazioni meno precise.

Attività: Creare un esempio simile e verificarlo con un LLM

Esperimento in aula: stesso prompt, oggi

Confrontiamo il comportamento storico con modelli contemporanei.

1. Copiare il prompt originale → 2. Eseguirlo con un modello attuale → 3. Confrontare gli output

Griglia di osservazione

- correttezza
- qualita della spiegazione
- uso di esempi
- gestione dell'incertezza
- rispetto del compito
- stile comunicativo

Domanda critica

Un output piu fluido e sempre anche piu affidabile?

Quali segnali usiamo per distinguere competenza apparente e qualita effettiva?

Obiettivo: non "scegliere il migliore", ma imparare a valutare un sistema AI.

Capacità Emergenti degli LLM: Generazione di Codice

Gli LLM possono generare codice in diversi linguaggi.

Prompt: Scrivi una funzione Python per il massimo comun divisore (GCD) usando la sottrazione.

GPT-4 ha prodotto codice corretto e leggibile. È stato in grado di riscrivere lo stesso codice in altri linguaggi.

python

Copy code

```python
def gcd_subtraction(a, b):
    while a != b:
        if a > b:
            a = a - b
        else:
            b = b - a
    return a
```

Gli LLM possono sostituire i programmatori? Quali sono i loro limiti?

Capacità Emergenti degli LLM: Creatività visiva

2023: descrizione testuale

Esempio: Disegno e Cartoni Animati

Prompt: "Descrivi un cartone in stile Gary Larson con mucche e alieni."

Risultato:

"Una mucca indossa una maschera aliena, mentre gli alieni, confusi, confrontano la mucca mascherata con l'immagine di una vera mucca."

Didascalia: "Mildred, non credo che ci stiano cascando."

Dalla descrizione alla generazione multimodale

2023: LLM testuale

Prompt testuale → descrizione testuale di un’immagine o di una scena.

Oggi: modello multimodale

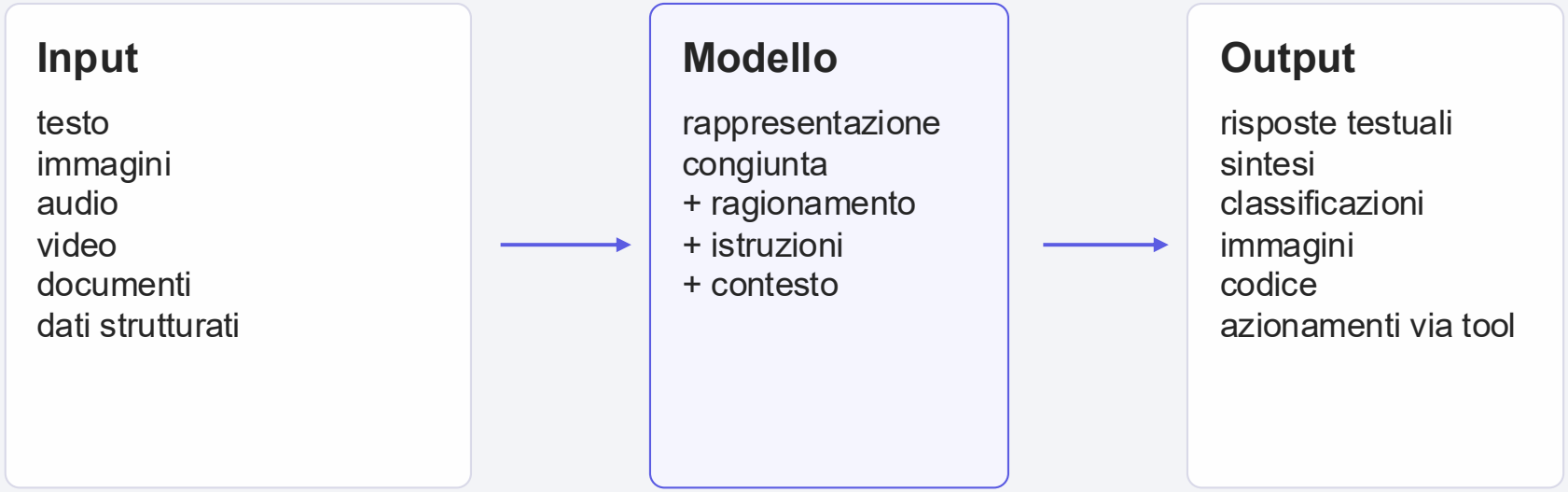
Prompt + immagini + documenti → testo, analisi visiva, generazione o modifica di contenuti.



Esempio generato: paesaggio naturalistico ad alta definizione

Multimodalita: input e output

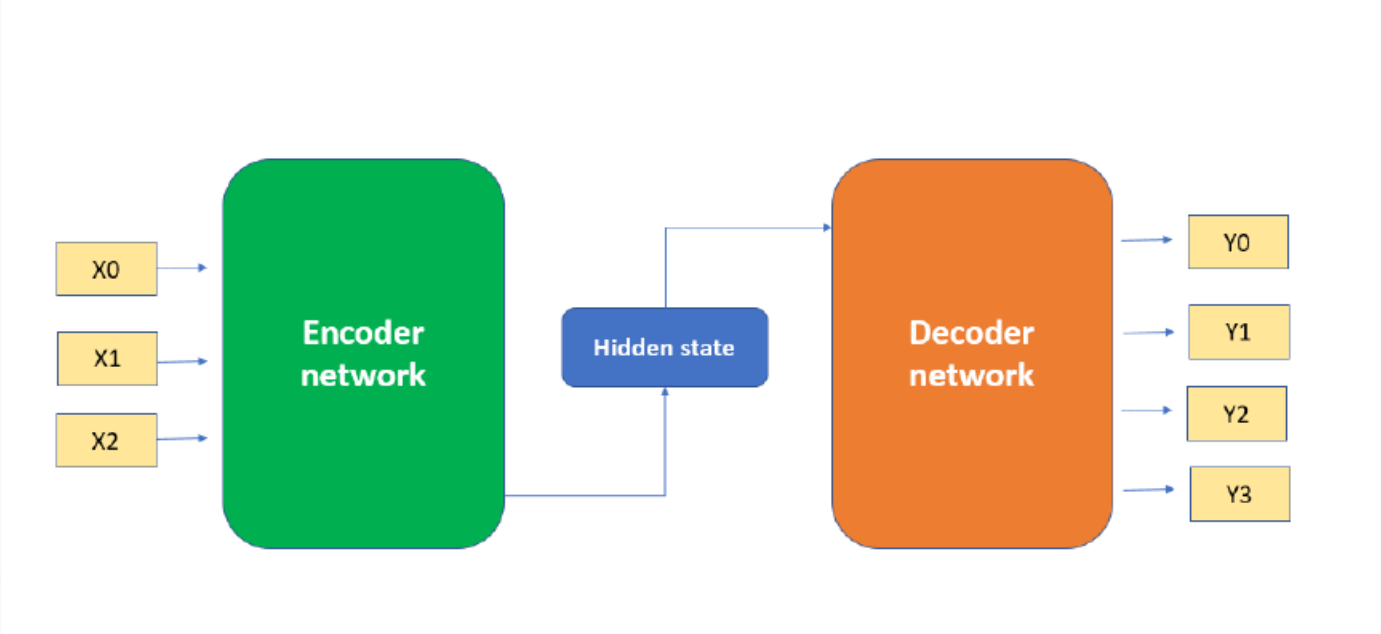
Il modello non lavora piu solo con sequenze di parole.



Per la comunicazione: analisi di fonti, briefing, creativita visuale, adattamento cross-media.

Come Funzionano gli LLM

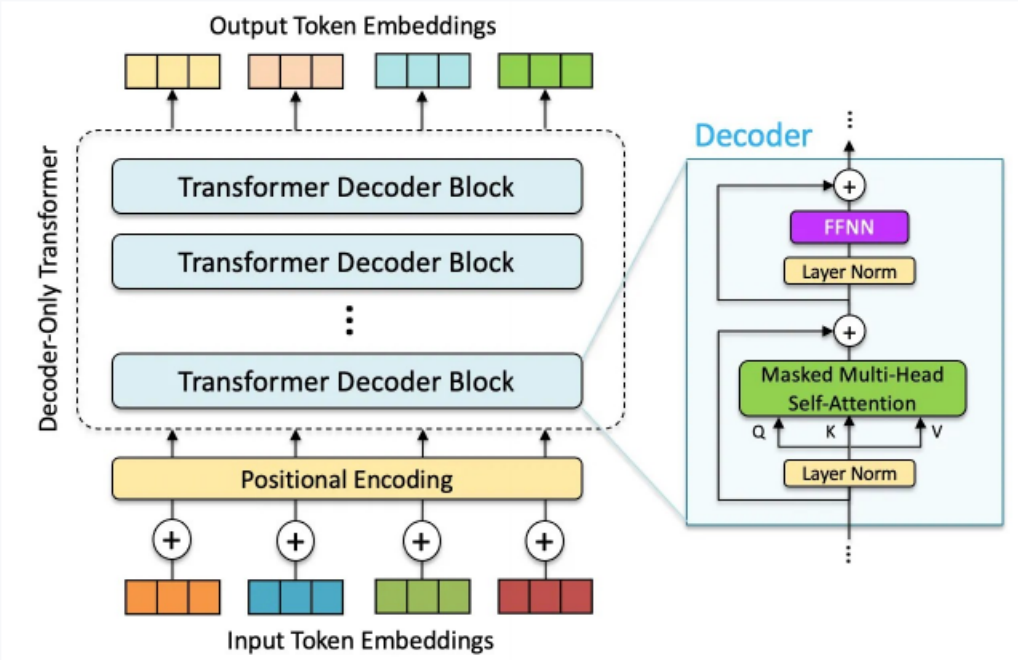
- L'Architettura del Trasformatore
  - Inventata nel 2017 con il paper "Attention Is All You Need".
  - Elementi principali:
    - » Encoder: Cattura le relazioni nei dati di input (non usato nei modelli GPT).
    - » Decoder: Genera il testo di output token per token.



Come Funzionano gli LLM (2)

- Meccanismo di Attenzione
  - Gli LLM assegnano un'importanza relativa ai token in una sequenza.
  - Questo consente di catturare relazioni semantiche anche a lunga distanza.

Decoder-only Transformer Architecture



Dal pre-training al post-training

1. Pre-training
previsione del token successivo su grandi raccolte di dati

2. Instruction tuning
adattamento a istruzioni, esempi e formati conversazionali

3. Alignment
feedback umano/AI e preferenze per ridurre risposte dannose

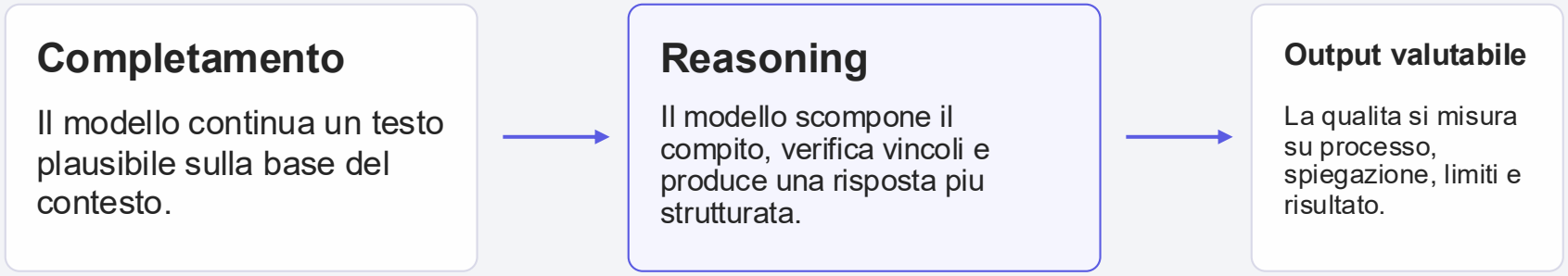
4. In-context learning
uso di prompt, esempi e documenti senza aggiornare i pesi

5. Inference-time reasoning
piu elaborazione per problemi complessi

6. Tool use
ricerca, codice, file, API e applicazioni esterne

Messaggio chiave: il “modello” e solo una parte del sistema AI.

Reasoning: cosa cambia rispetto al semplice completamento



Esempi didattici

problema logico con informazioni nascoste

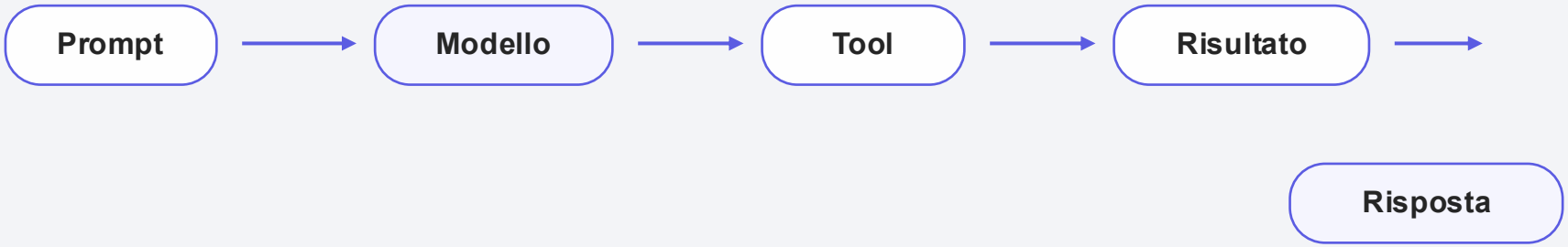
valutazione di una campagna di comunicazione rispetto a criteri espliciti

costruzione di una scaletta motivata per un comunicato o un piano editoriale

Importante: spiegazione convincente non equivale automaticamente a verita.

Tool use: quando il modello usa strumenti

Un LLM puo generare risposte; un sistema AI puo anche interrogare strumenti esterni.



Esempi di tool

ricerca web
file e documenti
code interpreter
calendario / email
fogli di calcolo
API aziendali

Per la comunicazione

analizzare rassegna stampa
creare report
aggiornare un piano editoriale
preparare bozze con fonti
richiedere approvazione umana

Da assistant a agent

La differenza non e solo “piu autonomia”: cambia il tipo di delega.

Assistant

risponde a domande e aiuta l’utente in conversazione

Esempio: “scrivi tre versioni di un post”

Workflow

esegue passi predefiniti in una sequenza

Esempio: form → sintesi → documento → email

Agent

riceve un obiettivo, pianifica passi, usa tool e puo chiedere conferme

Esempio: prepara un briefing settimanale da fonti diverse

Regola didattica: chiamare “agente” solo un sistema che combina obiettivo, stato, strumenti e ciclo di azione.

Limiti e sfide etiche: cosa resta attuale

Allucinazioni: risposte plausibili ma non verificate.

Bias: i modelli possono riflettere squilibri e stereotipi presenti nei dati o nei feedback.

Multimodalita: immagini, audio e video aumentano le opportunita ma anche i rischi di manipolazione.

Tool use: un errore puo trasformarsi in azione su file, email, database o workflow.

Agenti: servono permessi, tracciabilita, approvazione umana e limiti chiari.

Per la comunicazione: verificare fonti, contesto, pubblico, tono e responsabilita editoriale.

Human-in-the-loop: dove mettere il controllo umano

Automatizzare non significa eliminare la responsabilita professionale.

Basso rischio

brainstorming
varianti di titolo
riformulazioni
scalette

Controllo richiesto

sintesi di fonti
comunicati
contenuti social
analisi reputazionale

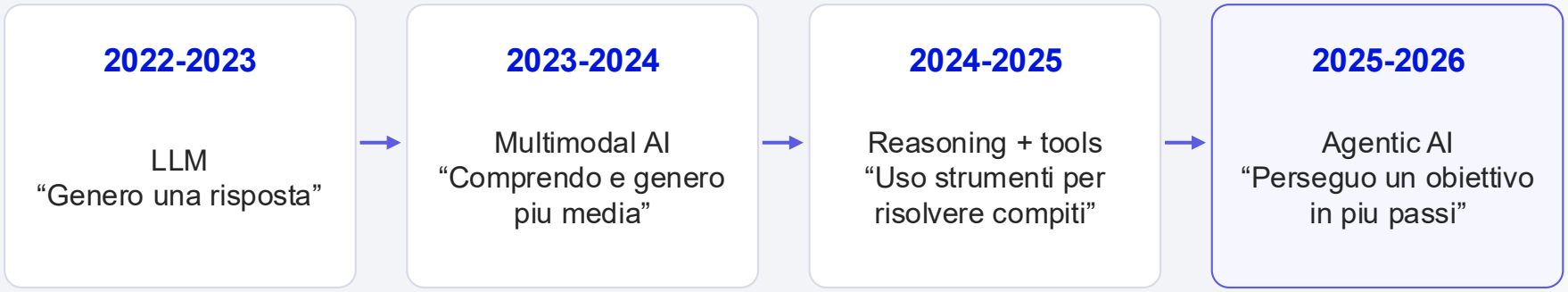
Alta attenzione

invio email
pubblicazione
uso dati personali
risposte istituzionali
contenuti sensibili

Checklist minima prima di pubblicare

- fonti verificate e citabili
- tono coerente con organizzazione e pubblico
- privacy, copyright e trasparenza valutati
- approvazione umana per azioni esterne
- log del processo e delle fonti usate
- piano di correzione se emergono errori

Da LLM a sistemi AI



Gli LLM restano il nucleo generativo, ma sempre piu spesso sono integrati in sistemi con memoria, fonti, strumenti e workflow.

Per la comunicazione, il punto centrale diventa progettare processi affidabili, non solo scrivere buoni prompt.

Principali Large Language Models oggi

Panorama aggiornato per il corso “IA per la comunicazione”

2026 | reasoning | multimodalità | tool use / agents

Obiettivo didattico: non “incoronare” un modello, ma capire famiglie, trade-off e casi d’uso.

Come leggere il panorama dei modelli

Un LLM oggi non è solo un generatore di testo: è spesso parte di un sistema più ampio.

1. Modello

Architettura, parametri, training, post-training, capacità di ragionamento.

2. Interfaccia

Chat, file, immagini, voce, canvas, integrazione con documenti e ricerca.

3. Strumenti

Web search, codice, browser, API, connettori, automazioni e computer use.

4. Deploy

Cloud, API, locale, open-weight, privacy, costi e licenze.

Per il corso conviene confrontare i modelli per "famiglie" e casi d'uso, non solo per nome commerciale.

- proprietari/frontier
- ecosistemi integrati
- open-weight e deploy controllato
- modelli specializzati per coding, voce, immagini, video

Modelli proprietari / frontier

Sono accessibili soprattutto tramite app e API; i dettagli interni sono spesso non pubblici.

OpenAI
GPT-6 family

Astra: modello flagship per ragionamento, coding e professional work.
6.1 Sol: equilibrio capacità/costo.
Luna: più veloce ed economico.

Google
Gemini 3

Modello multimodale orientato a reasoning, coding, tool use e integrazione con prodotti Google, AI Studio e Vertex AI.

Anthropic
Claude

Famiglia focalizzata su coding, knowledge work, uso quotidiano e task agentici a più passaggi.

xAI
Grok

Grok 4.7: modello flagship per codice, tool calling agentico e reasoning configurabile.

Messaggio per gli studenti: la “qualità” dipende dal compito, non solo dal nome del modello.

Modelli aperti e open-weight

Più trasparenza su dimensioni, licenze e deploy: utili per ricerca, didattica e progetti controllati.

Famiglia	Esempi attuali	Dimensioni pubbliche	Perché è interessante
Mistral	Mistral 3 / Large 3	675B totali, 41B attivi	Apache 2.0, multimodale, forte su multilingua
DeepSeek	DeepSeek V3.2	671B totali, 37B attivi	MoE, 128K context, licenza MIT
Qwen	Qwen3.6-27B	27B dense	open weights, multimodale, agentic coding
Meta	Llama ecosystem / Muse Spark	varia per modello	open innovation + modello multimodale in Meta AI

Questi modelli mostrano che “open” non significa necessariamente piccolo: spesso usano architetture MoE con pochi parametri attivi per token.

Dimensioni: cosa si sa e cosa no

Per molti modelli proprietari parametri e token di training non sono dichiarati pubblicamente.

Parametri

Valori numerici appresi dal modello.

Nei modelli aperti spesso sono dichiarati; nei modelli proprietari spesso no.

Token di training

Quantità di testo/elementi usati in addestramento.

Sempre meno spesso disponibile per i modelli frontier chiusi.

Context window

Quanti token il modello può usare in input/output durante una sessione.

È una misura più spesso pubblicata nelle API.

Regola didattica: non dire “X è migliore perché è più grande”. Dimensione, qualità dei dati, post-training, reasoning e tool use interagiscono.

Cosa confrontare davvero

Una griglia utile per scegliere quale modello mostrare in aula o usare in un progetto.

Dimensione	Domanda guida	Indicatori	Esempio di scelta
Reasoning	Serve risolvere problemi complessi?	effort, catena di passi, verifica	frontier reasoning model
Multimodalità	Servono immagini, audio, video?	input/output supportati	Gemini, Grok, GPT, Claude
Tool use	Deve agire su file, web o app?	API, web, browser, computer use	assistant/agent workflow
Deploy	Dove deve girare?	cloud, locale, licenza, costo	Mistral/Qwen/DeepSeek/Llama
Affidabilità	Quanto conta la verificabilità?	citazioni, grounding, log, human-in-loop	RAG + fonti + revisione

Esempi di uso didattico

Non serve mostrare tutti i modelli: basta scegliere 2–3 strumenti per far emergere le differenze.

Scrittura e comunicazione

Confronto di stile, tono, aderenza al brief e gestione delle fonti.

Analisi di documenti

Caricare fonti, chiedere sintesi, verificare citazioni e allucinazioni.

Coding / agentic tasks

Far vedere tool use, pianificazione, esecuzione di passi e verifica.

Open-weight e privacy

Discutere licenze, deploy locale, costi e controllo sui dati.

Suggerimento: una demo efficace usa lo stesso prompt su modelli diversi e valuta il risultato con una rubrica comune.

Mini-lab: confronto tra LLM oggi

Obiettivo: capire trade-off, non stabilire un vincitore universale.

Prompt comune

> *“Prepara un piano di comunicazione per promuovere un evento locale, indicando target, messaggi chiave, canali e rischi.”*

Criterio	Modello A	Modello B	Osservazioni
Aderenza al brief			
Qualità del ragionamento			
Uso di fonti / incertezza			
Stile comunicativo			
Utilità operativa			

Discussione finale: quale modello usereste per brainstorming? per analisi documentale? per workflow/agent? perché?