AI generativa per l'analisi dei percorsi clinici: Da Assistenti a Sistemi Agentici

Giuseppe Manco
Dirigente di Ricerca
Istituto di Calcolo e Reti ad Alte Prestazioni
Consiglio Nazionale delle Ricerche

L'anno scorso: GenAI per i referti medici

Architetture multimodali

- Vision Encoders + Text Decoders per referti da immagini DICOM
- Modelli specializzati: GPT-4V, LLaVA-Med, CLIP/BLIP/VLP
- Sistemi conversazionali per suppor

Opportunità identificate

- Generazione automatica di bozze di referti radiologici
- Didascalie cliniche e quantificazione lesioni
- visuale alla diagnosi differenziale

***Domanda chiave*: possiamo fidarci?**

Rischi tecnici rilevati

- Allucinazioni testuali e errori visivi sull'immagine
- Shortcut learning e bias da dataset (MIMIC-CXR, CheXpert)
- Distributional shift su immagini out-of-distribution

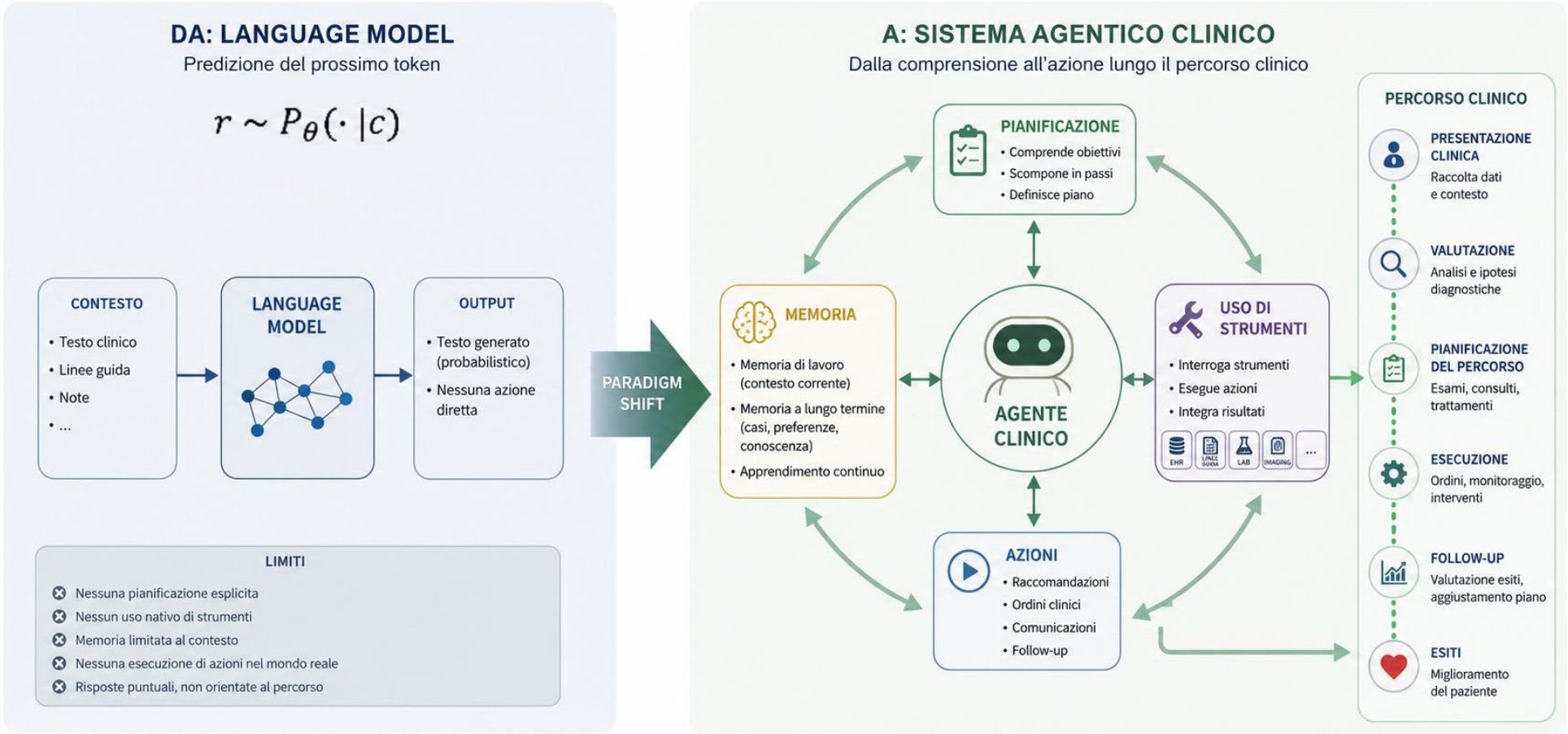
tivi ed etici

- EU AI Act: trasparenza obbligatoria, Grad-CAM / attention maps
- Validazione clinica FDA/EMA prima del deploy
- Responsabilità medico-legale ancora irrisolta

***In sintesi*:** *l'AI generativa multimodale è un "copilota clinico" promettente*

Sistemi agentici in medicina

Obiettivo:
Migliorare gli esiti del paziente



Che cosa cambia quando un LLM diventa agente

Un agente non produce solo una risposta: mantiene stato, sceglie il prossimo passo e può invocare strumenti.

Definizione operativa

LLM + orchestratore + stato + memoria + strumenti + policy di sicurezza

1. Osserva stato e contesto → 2. Ragiona e pianifica → 3. Agisce con tool/API → 4. Valuta esito e rischio → 5. Aggiorna memoria/stato

feedback loop: ogni azione modifica lo stato del sistema e condiziona il passo successivo

Autonomia vincolata
workflow, soglie, autorizzazioni e human-in-the-loop

Output azionabile
query, chiamate API, ordini, comunicazioni, escalation

Qualità di traiettoria
la sicurezza va misurata lungo sequenze multi-step

Architetture che implementano l'agency

Nella pratica l'agente è quasi sempre un'architettura di orchestrazione, non un singolo modello.

controllo deterministico → **autonomia adattiva**

Workflow / graph
nodi, transizioni, stato esplicito

- massima controllabilità e auditabilità

ReAct / tool loop
reason → act → observe

- interazione dinamica con tool e API

Planner–executor
decomposizione in sotto-task

- task lunghi e multi-step complessi

Multi-agent
supervisor, specialisti, peer review

- collaborazione e riduzione bias cognitivi

Stack architetturale ricorrente

Interfaccia / task → **Orchestratore state machine / graph** → **LLM ragionamento** → **Tool + dati API, EHR, RAG** → **Guardrail + audit HITL, policy, log**

Da applicazioni generali a percorsi clinici

1. Decision support e diagnosi

Ragionamento diagnostico, piani terapeutici, validazione prescrizioni, clinical trial reasoning.

2. Dati clinici e documentazione

EHR analytics, rischio di mortalità/readmission, scribing clinico, report patient-friendly.

3. Training e simulazione

Pazienti sintetici, simulazioni ospedaliere, OSCE, ambienti operatori e copilot chirurgici.

4. Ottimizzazione dei servizi

Educazione paziente, data collection, supporto amministrativo e automazione di task non diagnostici.

L'impatto più realistico nel breve periodo è nei processi a basso rischio clinico e alta ripetitività; l'uso diagnostico richiede validazione più rigorosa.

Perché i percorsi clinici sono un caso peculiare

Accesso triage → Valutazione clinica → Diagnosi multimodale → Piano terapeutico → Esecuzione azioni → Monitoraggio follow-up

Caratteristiche distintive

Longitudinalità
stato e memoria devono seguire il paziente tra episodi, reparti e follow-up.

Multimodalità
note, EHR, immagini, laboratorio, segnali e linee guida vanno fusi in un contesto unico.

Azionabilità
il sistema può generare raccomandazioni, ordini, comunicazioni o escalation.

Responsabilità distribuita
medici, infermieri, caregiver e paziente mantengono ruoli e diritti decisionali.

Il problema non è solo inferire una risposta corretta, ma coordinare decisioni, ruoli e azioni nel tempo.

Conseguenza architetturale: il "clinical pathway agent" deve essere un **orchestratore vincolato**, non un agente conversazionale generalista.

Reasoning models: cosa aggiungono davvero

Il reasoning è utile quando aumenta qualità decisionale, ma non coincide automaticamente con spiegabilità.

Scomposizione
trasforma un problema complesso in sotto-passi controllabili

Verifica con strumenti
interroga guideline, calculator, EHR, retrieval

Controfattuali
valuta alternative: cosa cambia se modifico X?

Secondo parere
genera ipotesi e critica il piano prima dell'azione

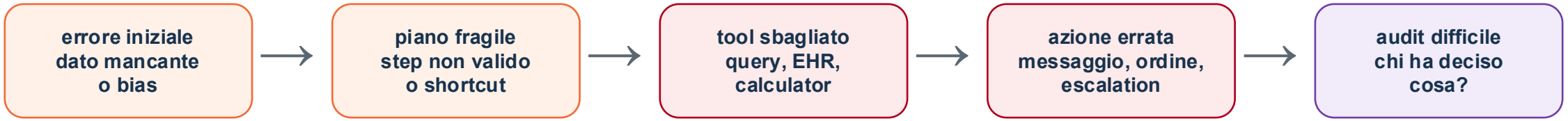
Trade-off da gestire

- più latenza e costo
- più punti di fallimento nella catena
- razionali plausibili ma non sempre fedeli
- calibrazione dell'incertezza ancora difficile

Servono evidenze, fonti, tool-call, controlli e criteri di escalation.

Quando l'agente sbaglia, il rischio si amplifica

Reasoning, memoria e tool-use aumentano la capacità del sistema, ma anche la superficie di fallimento.



I problemi aperti

C1 — Brittleness
distribution shift, casi rari, comorbidità, bias fuori training distribution

C2 — Self-correction debole
l'agente spesso scopre l'errore dopo l'azione

C3 — Opacità / Plausibility gap
decision trace plausibile ma non necessariamente fedele o auditabile

C4 — Coordination failures
specialisti agentici che non modellano capacità, obiettivi e limiti degli altri agenti

C5 — Safety non verificata
logging e guardrail non equivalgono a garanzie o evidenza clinica

Vincoli essenziali per il deployment clinico

Le capacità agentiche devono essere incapsulate in un "safety envelope" verificabile — tre principi non negoziabili.

1

Permissioning e autonomia graduata

- azioni cliniche separate da suggerimenti: scrivere, ordinare, escalare richiedono autorizzazione esplicita
- soglie di confidenza e rischio prima di ogni azione; human-in-the-loop proporzionale all'impatto paziente
- standard emergenti: MCP (Anthropic), A2A (Google) per interoperabilità multi-agente sicura

2

Osservabilità e audit trail

- ogni azione deve produrre una traccia leggibile: input, reasoning summary, tool call, fonte, outcome
- EU AI Act Annex III: clinical decision support = alto rischio → verifica formale
- log in formato auditabile per responsabilità professionale e post-market monitoring

3

Memoria sicura e privacy

- GDPR Art. 9: dati sanitari = categoria speciale; minimizzazione, consenso esplicito, diritto all'oblio
- federated learning per training su dati clinici senza centralizzazione
- separazione dei contesti: dati di un paziente non contaminano reasoning su un altro

Una risposta possibile: architettura a 3 livelli

Tier 1 - Reattivo

RAG + regole + tool semplici
risposte routinarie e alta confidenza

risolve C1 (brittleness) per i casi noti

Tier 2 - Deliberativo

reasoning causale, ipotesi controfattuali
controlli su evidenze e coerenza

risolve C2 (self-correction) + C3 (opacità)

Tier 3 - Metacognitivo

critic indipendente, safety check
monitoraggio, escalation e rollback

risolve C4 (coordinamento) + C5 (safety)

routing per confidenza + rischio + costo dell'azione

Componenti abilitanti

Provenance, coordinamento
fonti, dati, versioni

Formal checks
policy e invarianti

Human oversight
approvazione proporzionata

Audit trail
decision trace e post-market monitoring

Intuizione: Instradare ogni decisione al livello minimo sufficiente: veloce se routine, profondo se incerto, umano se critico

Validare un agente come sistema socio-tecnico

Dai benchmark statici alle traiettorie: azioni, recovery, impatto sul workflow, outcome e responsabilita

1 Simulazione
casi sintetici e stress test

→

2 Retrospettivo
dati storici, nessun impatto clinico

→

3 Shadow mode
raccomandazioni non visibili o non operative

→

4 Trial prospettico
endpoint di sicurezza, workflow e outcome

→

5 Deploy monitorato
drift, incidenti, rollback, change control

Trajectory metrics
task success, error recovery, unsafe action rate

Clinical metrics
appropriate escalation, time-to-decision, patient outcome

Operational metrics
clinician burden, latency, integration failures

Governance metrics
audit completeness, drift, incident response

La domanda cruciale è: non solo "quanto è accurato?", ma "che succede quando sbaglia?"

Tre cose da ricordare

1

I sistemi agentici con reasoning sono pronti per percorsi clinici a basso/medio rischio

— Agentico non significa autonomo: In sanità l'agente va pensato come orchestratore vincolato dentro workflow controllati.

2

Il reasoning migliora drammaticamente l'accuratezza

— ma introduce nuovi problemi di safety che i framework attuali non gestiscono

3

La Safety è una proprietà architetturale

— Routing a livelli, human oversight, audit trail e validazione progressiva, EU AI Act by design, standard aperti

Da copilota generativo a Sistema agentico affidabile: un salto tecnologico, progettuale, clinico e regolatorio.

Riferimenti

- Ambient Listening in Clinical Practice: Evaluating EPIC Signal Data Before and After Implementation and Its Impact on Physician Workload https://arxiv.org/abs/2504.13879
- AI Risk Management Framework | NIST https://www.nist.gov/itl/ai-risk-management-framework
- towards Generalist Biomedical AI https://arxiv.org/abs/2307.14334
- A Survey of LLM-based Agents in Medicine: How far are we from Baymax? https://arxiv.org/abs/2502.11211
- MedClarify: An information-seeking AI agent for medical diagnosis with case-specific follow-up questions https://arxiv.org/abs/2602.17308
- Towards Conversational Diagnostic AI https://arxiv.org/abs/2401.05654
- Large Language Models in Biomedical and Health Informatics: A Review with Bibliometric Analysis https://arxiv.org/abs/2403.16303
- A prospective clinical feasibility study of a conversational diagnostic AI in an ambulatory primary care clinic https://arxiv.org/abs/2603.08448
- Biomedical Large Languages Models Seem not to be Superior to Generalist Models on Unseen Medical Data https://arxiv.org/abs/2408.13833
- Good Machine Learning Practice for Medical Device Development: Guiding Principles | FDA https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles
- AI-Driven Healthcare: A Survey on Ensuring Fairness and Mitigating Bias https://arxiv.org/abs/2407.19655
- ReAct: Synergizing Reasoning and Acting in Language Models, https://arxiv.org/abs/2210.03629
- Reflexion: Language Agents with Verbal Reinforcement Learning https://arxiv.org/abs/2303.11366
- AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation https://arxiv.org/abs/2308.08155
- LangGraph docs, https://www.langchain.com/langgraph
- OpenAI Agents SDK docs, https://developers.openai.com/api/docs/guides/agents
- Google ADK docs, https://adk.dev/
- CrewAI docs, https://docs.crewai.com/en