UNA SERIE SULL'INTELLIGENZA ARTIFICIALE

Io, Gemini e altri quattro

Parte X

Il Progetto A.D.A.M.: Un'IA Personale per la Ricerca Ingegneristica

ROCCO MASSIMILIANO MONDO

ABSTRACT

Siamo arrivati all'ultimo episodio. Nove mesi fa siamo partiti da Las Vegas, dal CES 2026, con la promessa che la rivoluzione dell'intelligenza artificiale non fosse una questione per soli addetti ai lavori. Oggi chiudiamo il cerchio con un progetto concreto, completamente funzionante, costruito mattone su mattone attraverso tutto ciò che abbiamo esplorato in questa serie. La Parte IX ci ha consegnato l'anima di A.D.A.M., la sua architettura software, la sua forma-mentis STEM, i suoi layer epistemici. Questa Parte X gli darà un corpo. Parleremo dell'hardware che ospita il sistema, del difficile equilibrio tra prestazioni e costo che affronta ogni ricercatore indipendente, e di come una scelta tecnologica scalabile possa trasformarsi nel tempo in qualcosa di ben più potente. Poi vedremo A.D.A.M. al lavoro: lo stato reale del sistema, il corpus che sta crescendo, le prime applicazioni concrete nella ricerca scientifica. Infine alzeremo lo sguardo verso l'orizzonte: l'intelligenza artificiale generale l'AGI – non è più fantascienza. È un appuntamento con la storia che si avvicina, con tutto ciò che questo comporta. Alla fine, un paragrafo per chi ha avuto la pazienza di arrivare fino in fondo.

Rocco M. Mondo, 10 ottobre 2026

PARTE X

IL PROGETTO DEFINITIVO: COSTRUIRE UN CORPO PER A.D.A.M.

PREMESSA

Nella Parte IX vi ho lasciato con una promessa: «un'anima non può esistere sul piano materiale se non ha anche un corpo». Quella promessa oggi si compie. Ma compiere una promessa in campo tecnologico, soprattutto quando si è un ricercatore indipendente, senza finanziamenti esterni, senza un laboratorio universitario alle spalle, significa fare i conti con una realtà molto concreta: il denaro. E prima ancora del denaro, la strategia.

Questo ultimo episodio è, in un certo senso, il più onesto dell'intera serie. Non parleremo solo di architetture eleganti e visioni future. Parleremo di scelte difficili, di compromessi necessari, di come si costruisce qualcosa di reale con risorse limitate. E di come, a volte, la scelta obbligata si rivela, con il senno di poi, anche la scelta migliore.

PARTE I: LA SCELTA DELL'HARDWARE (IL CORPO DI A.D.A.M.)

Il vincolo del ricercatore indipendente

Chi lavora all'interno di un'istituzione accademica o di un'azienda tecnologica ha accesso a risorse che la maggior parte delle persone non può nemmeno immaginare: cluster GPU con centinaia di schede NVIDIA A100, budget di training da decine di migliaia di euro al mese, team di ingegneri dedicati. Il ricercatore indipendente no. Ha un budget, una stanza, e se è fortunato, una mente capace di trasformare i vincoli in creatività.

Il mio vincolo principale era chiaro sin dall'inizio: il sistema doveva essere locale, doveva ospitare modelli da 70 miliardi di parametri in modo fluido, e doveva essere acquistabile con un budget ragionevole. Non esisteva, all'inizio del 2026, una risposta ovvia a questa equazione. Esisteva però un'architettura emergente che stava cambiando le regole del gioco.

AMD Ryzen AI Max+ 395 «Strix Halo»: la piattaforma di A.D.A.M.

La scelta è caduta su un processore che, nel momento in cui scrivo, rappresenta probabilmente il caso più interessante nel mercato consumer per l'inferenza AI locale: l'AMD Ryzen AI Max+ 395, nome in codice «Strix Halo». Non è una GPU dedicata. Non è un server. È un System-on-Chip

progettato per i laptop e i mini-PC di fascia alta, con una caratteristica che lo rende unico nel panorama AI: la memoria unificata.

Caratteristica	AMD Ryzen AI Max+ 395 «Strix Halo»
Architettura CPU	16 core Zen 5 / 32 thread — fino a 5.1 GHz boost
GPU integrata	AMD Radeon 8060S — 40 Compute Units RDNA 3.5
NPU dedicata	AMD XDNA 2 — 50 TOPS per inferenza AI leggera
Memoria unificata	128 GB LPDDR5X-8000 (96 GB disponibili come VRAM)
Banda di memoria	256 GB/s — il vero collo di bottiglia dell'inferenza LLM
Architettura	UMA — CPU e GPU condividono fisicamente la stessa RAM
Consumo tipico AI	45-120W TDP configurabile — efficiente per uso continuativo

Il vantaggio fondamentale di questa architettura UMA (Unified Memory Architecture) per il caso d'uso specifico di A.D.A.M. è geometrico. Un modello LLM quantizzato a Q4_K_M da 70 miliardi di parametri occupa circa 40 GB in memoria. Su un sistema tradizionale con GPU dedicata questo richiederebbe una scheda con almeno 40 GB di VRAM — praticamente irraggiungibile nel mercato consumer. Su Strix Halo, i 128 GB di RAM fisica sono disponibili sia per la CPU che per la GPU senza alcun trasferimento su bus PCIe. Il modello si carica, ci vive, e la GPU lo accelera direttamente. Restano ancora 88 GB per il sistema operativo, per il backend FastAPI, per il context esteso da 256k, per la pipeline completa di A.D.A.M. Plus.

Il confronto: NVIDIA, Apple e la realtà del mercato

Sarebbe disonesto presentare Strix Halo come l'unica scelta possibile senza confrontarlo con le alternative reali. Vediamole.

Piattaforma	Max VRAM/RAM AI	Prezzo indicativo	Limitazioni per A.D.A.M.
AMD Strix Halo (128GB)	96 GB condivisa	€1.500–2.500 (mini-PC)	Ecosistema ROCm meno maturo di CUDA
NVIDIA RTX 4090 (desktop)	24 GB VRAM	€1.800–2.200 (solo GPU)	24 GB troppo limitati per 70B Q4; richiede sistema completo
NVIDIA A100 80GB (server)	80 GB HBM2e	€10.000–15.000+	Fuori budget per ricercatore indipendente; consumi elevati
Apple M3 Ultra (192GB)	192 GB condivisa	€8.000–12.000 (Mac	Ecosistema chiuso; costo proibitivo;

		Studio/Pro)	aggiornamento difficile
Apple M4 Pro (64GB)	64 GB condivisa	€3.000–4.500 (MacBook Pro)	Ecosistema chiuso; meno flessibile; ROCm non disponibile
NVIDIA RTX 5090 (2025)	32 GB GDDR7	€2.500–3.000 (solo GPU)	Migliora su 4090 ma ancora insufficiente per 70B pieno

Il quadro è chiaro. Apple offre la soluzione tecnicamente più elegante con i suoi 192 GB su M3 Ultra — ma a un costo che esclude quasi completamente il ricercatore indipendente, e all'interno di un ecosistema chiuso che rende difficile l'integrazione con i tool standard dell'AI open-source. NVIDIA domina il mercato professionale, ma il divario tra la fascia consumer (24 GB su RTX 4090) e quella professionale (80 GB su A100) è abissale in termini di costo. Strix Halo occupa un posizionamento unico: prestazioni sufficienti per i modelli target di A.D.A.M., prezzo accessibile, ecosistema Linux/Windows maturo.

Sul fronte software, l'ecosistema ROCm/HIP di AMD — per quanto meno maturo di CUDA — ha fatto passi da gigante tra il 2025 e il 2026. Il supporto in llama.cpp e llama-cpp-python per l'accelerazione GPU AMD è solido e la community è attiva. Non è CUDA, ma è sufficiente per i nostri scopi.

Scalabilità: il cluster a quattro nodi come orizzonte

Uno degli aspetti che più mi ha convinto nella scelta di Strix Halo non è la configurazione attuale, ma quella futura. La piattaforma espone due interfacce che trasformano un singolo nodo in un potenziale elemento di un cluster distribuito.

USB4 v2 a 80 Gbps: la seconda generazione di USB4 raddoppia la banda disponibile rispetto alla prima (da 40 a 80 Gbps). Non è una porta per la ricarica del telefono, è un'interfaccia capace di connettere due nodi Strix Halo con una larghezza di banda sufficiente per il trasferimento di attivazioni e pesi tra macchine. In un'architettura multi-nodo per l'inferenza distribuita, questo può permettere di aumentare la memoria RAM disponibile sommando quella di più sistemi.

OCuLink: ancora più interessante per le applicazioni AI. OCuLink è essenzialmente PCIe esternalizzato, permette di collegare GPU esterne (eGPU) con una larghezza di banda prossima

a quella di uno slot PCIe x4 interno (~32 Gbps). Questo apre scenari concreti: collegare una GPU NVIDIA dedicata (come una RTX 4090 o una RTX 5090) al sistema Strix Halo, ottenendo il meglio dei due mondi, la RAM unificata di AMD per i modelli grandi, la CUDA-acceleration di NVIDIA per operazioni specifiche che la supportano meglio.

L'architettura a cluster immaginata per A.D.A.M. prevede la possibilità di collegare fino a quattro nodi Strix Halo tramite queste interfacce, gestiti da software come llama.cpp in modalità server distribuito o framework come LM Distributed. Il risultato teorico: 4 × 96 GB = 384 GB di memoria AI distribuita, a una frazione del costo di un singolo A100 da 80 GB. Non è ancora implementato ma è l'orizzonte verso cui il progetto può crescere senza cambiare piattaforma.

[Note sull'ecosistema software]

L'importanza di un ambiente software maturo non va sottovalutata. Scegliere hardware performante con un ecosistema software immaturo significa perdere la metà delle prestazioni in configurazione e debug invece che in ricerca. ROCm 7.x, Python 3.12, llama-cpp-python con supporto HIP, FastAPI, tutto questo forma uno stack che funziona oggi, in produzione, su questa piattaforma. È un fattore di scelta tanto importante quanto i gigabyte di RAM.

PARTE X

IL PROGETTO DEFINITIVO: COSTRUIRE UN CORPO PER A.D.A.M.

PARTE II: DA PROTOTIPO A SISTEMA OPERANTE

La transizione più difficile

Nella storia di qualsiasi progetto tecnologico, esiste un momento critico che molti non nominano perché è scomodo: il passaggio dal prototipo che «funziona come demo» al sistema che funziona nella quotidianità. Un prototipo è per definizione un atto di ottimismo, mostra cosa è possibile, non cosa è affidabile. Un sistema dichiarato "operativo" è un atto di ingegneria, deve funzionare quando non ci sei tu a supervisionarlo, su input che non hai previsto, in condizioni che non hai simulato.

A.D.A.M. ha attraversato questa transizione. Non è stato un momento preciso è stato un processo di mesi, fatto di test falliti che hanno insegnato qualcosa, di configurazioni ottimizzate e poi ri-ottimizzate, di pipeline che sembravano completate e poi si rivelavano fragili su casi limite. Il risultato di questo percorso è un sistema che oggi non definisco più «prototipo».

Il corpus: la conoscenza che cresce

La componente più viva di A.D.A.M. Plus — quella che continua a crescere giorno dopo giorno è il corpus di conoscenza. È inutile descriverne i contenuti specifici in questa sede: parte del valore di un sistema di ricerca è nella riservatezza delle fonti e delle direzioni di indagine. Quello che posso descrivere è la struttura e il processo.

Il corpus si articola su due livelli distinti. Il primo livello è la conoscenza strutturata di base: testi scientifici fondamentali, normative tecniche, letteratura di riferimento nel dominio STEM di interesse. Questo livello cresce lentamente ma solidamente, ogni documento ingerito viene processato dalla pipeline LLM Wiki, scomposto in concetti atomici, collegato agli altri da relazioni semantiche esplicite, e reso navigabile attraverso Obsidian come se fosse un grafo di pensieri. Il risultato non è un database di testi: è una rete di conoscenza che A.D.A.M. può percorrere come un essere umano percorre la propria memoria associativa.

Il secondo livello è la conoscenza contestuale: i risultati intermedi del lavoro di ricerca, le

ipotesi formulate e i loro status di verifica, le connessioni emergenti tra domini apparentemente lontani. Questo livello è per sua natura più volatile e più prezioso. È qui che A.D.A.M. diventa qualcosa di più di un sistema RAG sofisticato: diventa un registro vivente del pensiero scientifico in corso.

Le applicazioni concrete nella ricerca scientifica

La domanda più importante per qualsiasi sistema AI dedicato alla ricerca non è «cosa può fare in teoria?» ma «cosa fa concretamente, oggi, per migliorare il lavoro?». Per A.D.A.M., la risposta si articola in tre categorie di utilizzo che si sono consolidate nel tempo.

Analisi di ipotesi complesse e verifica epistemica. Il layer epistemico di A.D.A.M. Plus ha dimostrato il suo valore più chiaramente in questa funzione. Sottoporre un'ipotesi di ricerca al sistema, un'ipotesi di confine che sfida i modelli consolidati, produce un'analisi articolata in quattro dimensioni: coerenza interna, compatibilità con i principi fisici noti, fattibilità sperimentale con identificazione del Minimum Viable Experiment, e argomentazione contraria sistematica. Non è una risposta. È una conversazione strutturata che forza la chiarezza del pensiero.

Correlazione cross-dominio. Uno dei limiti più frustanti della ricerca specialistica è la difficoltà di tenere simultaneamente in vista sviluppi in domini contigui ma formalmente separati. Un fenomeno della meccanica quantistica che ha implicazioni nell'ingegneria dei materiali. Una tecnica di commutazione di potenza che ha analogie con un processo di transizione di fase. A.D.A.M., grazie alla struttura a grafo del suo corpus, è in grado di suggerire connessioni di questo tipo non come certezze, ma come piste da esplorare.

Generazione e verifica di documentazione tecnica. La produzione di relazioni tecniche, la verifica della coerenza interna di un documento complesso, la generazione di sezioni strutturate a partire da note di lavoro non ordinate: questi compiti che consumano tempo

prezioso possono essere delegati ad A.D.A.M. con un livello di affidabilità che si è consolidato

nel tempo. Il layer epistemico garantisce che le affermazioni normative siano supportate dalle fonti presenti nel corpus, e segnala esplicitamente quando si va oltre i confini della conoscenza verificabile.

Quello che accomuna queste tre applicazioni è un principio che ho imparato a riconoscere nel tempo: A.D.A.M. non sostituisce il pensiero MA lo amplifica. La qualità dell'output dipende sempre e inevitabilmente dalla qualità dell'input, dalla chiarezza della domanda, dalla precisione del contesto fornito. Un sistema AI è tanto migliore quanto migliore è il ricercatore che lo usa. Questa non è una limitazione: è la sua natura più profonda.

PARTE III: L'AGI È ALLE PORTE

Cosa sta succedendo davvero

Nel gennaio 2026, quando ho scritto la Parte I di questa serie ispirandomi al CES di Las Vegas, il panorama dell'intelligenza artificiale era già in rapida evoluzione. In meno di un anno, la velocità di quel cambiamento è aumentata ulteriormente. Non in modo lineare ma in modo esponenziale, con sviluppi che sarebbero sembrati esagerati anche sei mesi fa.

I modelli di ragionamento, le architetture che pensano prima di rispondere, che mostrano i passaggi del loro ragionamento in modo verificabile hanno fatto passi da gigante. I sistemi che prima allucinavono con fiducia ora dichiarano esplicitamente l'incertezza. I benchmark che sembravano irraggiungibili sono state superate. In sintesi, la domanda che un anno fa era ancora confinata nei paper accademici è diventata una domanda da prima pagina: quando arriverà l'AGI?

AGI: definizione onesta di un termine sovraccarico

Prima di rispondere alla domanda, vale la pena chiarire di cosa stiamo parlando. AGI (***Artificial General Intelligence)*** è un termine che significa cose diverse a seconda di chi lo usa. Per alcuni ricercatori, AGI è il momento in cui un sistema artificiale supera le capacità cognitive umane

in ogni dominio misurevole. Per altri, è il punto in cui un sistema può imparare qualsiasi nuovo compito con la stessa efficienza di un essere umano, senza essere specificamente addestrato su di esso. Per altri ancora, è semplicemente il momento in cui un sistema AI può operare autonomamente in modo economicamente significativo in una gamma sufficientemente ampia di contesti.

Le definizioni differiscono perché l'intelligenza umana stessa è difficile da definire. Quello che sappiamo è che i sistemi attuali, inclusi i più avanzati, sono ancora fondamentalmente sistemi di "pattern di riconoscimento" sofisticatissimi, non sistemi che comprendono nel senso profondo che attribuiamo a questo termine quando parliamo di esseri umani. La differenza tra «simulare la comprensione» e «comprendere» è filosoficamente profonda e praticamente irrilevante per molte applicazioni ma è tutto tranne che irrilevante per applicazioni che richiedono giudizio in situazioni del tutto nuove.

I segnali concreti di un cambio di paradigma

Al di là delle definizioni, quello che possiamo osservare sono i segnali concreti che il paradigma sta cambiando, non tra un decennio, ma nel giro di pochi anni.

Il ragionamento esteso e verificabile. I modelli con Extended Thinking, capaci di «pensare ad alta voce» prima di rispondere, hanno dimostrato capacità di problem-solving su problemi matematici e scientifici che fino a due anni fa erano prerogativa esclusiva di esperti umani. Non è solo performance: è trasparenza del processo. Poter seguire il ragionamento di un sistema AI è un prerequisito per fidarsi di quel ragionamento in contesti ad alta posta.

Il calcolo quantistico come moltiplicatore. Il chip Willow di Google Quantum AI (105 qubit fisici, errori che diminuiscono all'aumentare della scala) ha dimostrato nel 2024 qualcosa di storicamente significativo: la correzione degli errori quantistici scala favorevolmente. Questo

non significa che i computer quantistici sostituiranno i sistemi AI classici domani. Significa che l'ostacolo principale alla loro praticità, l'accumulo di errori che rendeva inutile aumentare la scala, ha trovato una risposta sperimentale. Per la ricerca in fisica quantistica applicata,

esattamente il dominio di A.D.A.M., questo è un segnale da monitorare con attenzione.

Cosa significa per la ricerca indipendente

L'arrivo dell'AGI, qualsiasi forma concreta possa assumere, cambierà il panorama della ricerca scientifica in modi difficili da prevedere con precisione. Quello che mi sembra ragionevole anticipare è questo: il vantaggio competitivo si sposterà dalla capacità di elaborare informazioni alla capacità di porre le domande giuste. Un sistema AGI potrà leggere tutta la letteratura scientifica pubblicata e trovare connessioni che nessun essere umano avrebbe mai visto. Ma le domande che vale la pena porre, le direzioni di ricerca che portano a scoperte nuove, nasceranno ancora dall'intuizione umana, dall'esperienza pratica, dalla capacità di vedere nella realtà ciò che i dati non mostrano ancora.

Questo è esattamente il ruolo per cui A.D.A.M. è stato progettato: non sostituire il ricercatore, ma liberarlo dai compiti cognitivi che consumano energia senza generare "***insight***". La verifica normativa, la sintesi della letteratura, la coerenza interna di un ragionamento complesso, tutto questo può essere delegato. Il salto concettuale, la domanda che nessuno aveva ancora posto, la connessione tra domini apparentemente lontani: questo resta umano. Per ora.

CONCLUSIONE DELLA SERIE

«Io, Gemini e altri quattro» nasce in un gennaio del 2026 come una serie su tutto ciò che stava accadendo nel mondo dell'intelligenza artificiale. Finisce, dieci episodi dopo, come qualcosa di più personale: il racconto di un percorso, tecnico, intellettuale, a volte frustrante, spesso sorprendente, che ha trasformato un ricercatore libero in qualcuno che costruisce sistemi AI per la ricerca di confine. Ho cercato in ogni episodio di mantenere un equilibrio difficile: essere abbastanza tecnico da essere utile a chi conosce il campo, abbastanza chiaro da non perdere chi si avvicina per la prima volta. Non so quanto ci sono andato vicino. So che ci ho provato con onestà.

I dieci episodi hanno coperto un terreno enorme: la storia dell'intelligenza artificiale da ***Turing*** a oggi (Parte II), l'anatomia interna dei modelli linguistici con i Transformer e il meccanismo di Self-Attention (Parte III), l'arte del prompt engineering con i suoi pattern e le sue tecniche

(Parte IV), il confronto tra i cinque grandi LLM del mercato (Parte V), gli agenti AI che agiscono nel mondo reale (Parte VI), il RAG e la memoria contestuale (Parte VII), l'applicazione pratica nell'ingegneria elettrica (Parte VIII), e infine la costruzione di A.D.A.M.: prima l'anima software (Parte IX) e poi il corpo hardware (questa Parte X).

UNA NOTA FINALE

C'è qualcosa di strano e bello nel costruire un'intelligenza artificiale quando si è un ricercatore indipendente, senza finanziamenti, senza un team, con una sola stanza e il ronzio discreto di un mini-PC che lavora anche mentre dormi.

Strano perché questa era, fino a pochissimi anni fa, l'esclusiva assoluta di aziende multimiliardarie e laboratori universitari con decenni di storia. Bello perché questa è la promessa che la serie ha cercato di mantenere fin dal primo episodio: la ***democratizzazione tecnologica*** non è uno slogan. È qualcosa che si può toccare con mano, costruire pezzo per pezzo, testare e migliorare in una stanza qualsiasi con strumenti open-source e hardware consumer.

A.D.A.M. non è un prodotto. È un argomento. Non dimostra che un singolo ricercatore può fare ciò che fa Google — sarebbe ridicolo affermarlo. Dimostra che il tipo di pensiero che un sistema AI può amplificare, il ragionamento tecnico di confine, la connessione tra domini lontani, la verifica epistemica delle ipotesi, non è più una prerogativa di chi ha risorse illimitate. È disponibile, con pazienza e metodo, anche per chi lavora da solo.

Dante *scriveva: «Fatti non foste a viver come bruti, ma per seguir virtute e canoscenza». Settecento anni dopo, la canoscenza ha preso forme che Dante non avrebbe potuto immaginare. Ma il principio regge: l'intelligenza, umana o artificiale, vale quanto la domanda che si pone, non quanto la risposta che produce.* **La domanda giusta è ancora e sempre umana.**

Rocco M. Mondo, 10 ottobre 2026