PILOT 2.000 STUDENTI

Piano di Progettazione e Implementazione

Budget Completo: Infrastruttura + Team + AI Tooling

5-6 Corsi • Architettura OIOD Rev 2.0 • 4 Strati

Moduli: DOC + Tutor + Socrates + Quiz + Simulazione Esame

Rev. 2 — Febbraio 2026

Synthetic Data S.r.l. — Confidenziale

GLOSSARIO

Questo glossario raccoglie tutte le abbreviazioni, gli acronimi e i termini tecnici utilizzati nel documento e nel progetto QWIKKEN.

Acronimi di Progetto

Sigla	Significato	Descrizione
QWIKKEN	Quick Knowledge Engine	Piattaforma di tutoring AI per università telematiche
OIOD	Offline Intelligence, Online Delivery	Architettura che pre-genera artefatti didattici offline e li serve in tempo reale con modelli leggeri
MMS	Meta Master System	Il sistema operativo cognitivo alla base di QWIKKEN (motore EGEMON)
EGEMON	Enterprise Genome Operational Nexus	Piattaforma AI proprietaria di Synthetic Data su cui si basa QWIKKEN
DCA	Deep Contextual Awakening	Profilazione cognitiva dello studente: stile di apprendimento, livello, preferenze
UC	Unità Concettuale	Atomo di conoscenza: un singolo concetto didattico mappato nel Knowledge Graph
PKL	Personal Knowledge Library	Libreria personale di conoscenza di ogni studente
SENSORIUM	—	Modulo di ingestione contenuti: parsing documenti, trascrizione video, estrazione metadati
SYNAPSE	—	Modulo di gestione della conoscenza: Knowledge Graph, mapping concetti-syllabus, gap analysis
SOCRATES	—	Modulo di tutoring maieutico: guida lo studente con domande progressive
ARENA	—	Modulo di simulazione e valutazione: quiz, esami simulati, previsione voto
SENTINEL	—	Modulo anti-ghostwriting: verifica che gli elaborati siano originali dello studente
CORTEX	—	Intent Translation Engine: comprende l’intenzione dell’utente e la converte in query eseguibili
CALCULATOR	—	Axiomatic Kernel: motore logico-matematico deterministico, zero allucinazioni

Acronimi Tecnici e di Business

Sigla	Significato	Contesto
LLM	Large Language Model	Modello AI di grandi dimensioni (es. Gemini, Claude)
COGS	Cost of Goods Sold	Costo diretto per erogare il servizio a uno studente
RAL	Retribuzione Annua Lorda	Stipendio lordo annuale di un dipendente
FTE	Full-Time Equivalent	Unità di misura dell'impegno lavorativo (1 FTE = 1 persona full-time)
DAU / WAU / MAU	Daily / Weekly / Monthly Active Users	Metriche di adozione del prodotto
NPS	Net Promoter Score	Indice di soddisfazione e raccomandazione (-100 a +100)
QPM / RPM / RPD	Query / Requests Per Minute / Day	Metriche di carico e rate limiting delle API LLM
TPM	Tokens Per Minute	Limite di throughput delle API LLM
P95	95° Percentile	Metrica di latenza: il 95% delle richieste completa entro questo tempo
SLA	Service Level Agreement	Garanzia contrattuale di uptime e performance
SSO	Single Sign-On	Autenticazione unificata (lo studente accede con le credenziali dell'ateneo)
MCP	Model Context Protocol	Protocollo Anthropic per connettere agenti AI a servizi esterni
RAG	Retrieval Augmented Generation	Tecnica che arricchisce le risposte LLM con documenti specifici
HA	High Availability	Architettura con ridondanza per garantire continuità di servizio
CEPO	Chief Executive / Partnership Officer	Dirigente responsabile delle partnership strategiche nell'ateneo
CdA	Consiglio di Amministrazione	Organo decisionale dell'ateneo
B2B2C	Business to Business to Consumer	Modello: Synthetic Data vende all'ateneo, che offre il servizio allo studente
TCO	Total Cost of Ownership	Costo totale di possesso di una soluzione nel tempo
ROI	Return on Investment	Rapporto tra guadagno e investimento

Termini Architettura OIOD

Termine	Definizione
La Fabbrica (Strato 1)	Pipeline batch offline che genera artefatti didattici usando modelli premium (Gemini 2.5 Pro). Opera una tantum per corso, non in tempo reale.
Il Bibliotecario (Strato 2)	Modello leggero (Gemini 2.0 Flash) che serve artefatti pre-generati in tempo reale. Gestisce il 75-80% delle interazioni a costo quasi zero.
L'Assistente Senior (Strato 3)	Modello intermedio (Gemini 3 Flash) per escalation di media complessità. Introdotto nella Rev 2.0 come buffer tra Bibliotecario e Professore.
Il Professore (Strato 4)	Modello premium (Gemini 3 Pro) per casi complessi: SOCRATES profondo, cross-concept, generazione esami. Solo 3-5% del traffico.
Artefatto didattico	Unità di contenuto pre-generata dalla Fabbrica: spiegazione, analogia, esempio numerico, quiz, albero socratico, FAQ.
Learning Loop	Ciclo virtuoso: le risposte migliori dell'Assistente Senior e del Professore vengono promosse ad artefatti, riducendo le escalation future.
Context Caching	Funzione Gemini che mette in cache il prefix di contesto (syllabus, system prompt), riducendo i token di input del 90%.
Semantic Cache	Cache custom su pgvector che intercetta domande simili e serve risposte già generate, riducendo le chiamate LLM del 30-40%.
Escalation	Quando il Bibliotecario non ha un artefatto adeguato, passa la richiesta allo strato superiore (Assistente Senior o Professore).
Cowork	Agente AI desktop di Anthropic (Claude Opus 4.6) che automatizza task multi-step: gestione file, creazione documenti, automazione workflow. Usato nel pilot come moltiplicatore di produttività del team di sviluppo.
Claude Code	Tool a riga di comando di Anthropic per coding agentico. Precursore tecnico di Cowork, usato dagli sviluppatori.

1. PERIMETRO DEL PILOT

Parametro	Valore	Note
Studenti	2.000	Iscritti attivi a 5-6 corsi
Corsi	5-6	Un semestre tipo, discipline diverse
UC stimate	~1.000-1.200	200 UC/corso × 5-6 corsi
Ateneo	Da definire	Target: università telematica italiana
Durata pilot	4 mesi (16 settimane)	Inclusi onboarding e valutazione
Architettura	OIOD Rev 2.0 — 4 strati	Fabbrica / Bibliotecario / Assistente Senior / Professore
Piattaforma base	QWIKKEN BASE Rev 5.0	Frontend esistente (React su Google AI Studio)

1.1 Moduli Attivi

Modulo	Priorità	Funzione	Strato OIOD
DOC (Syllabus)	P0	Ingestione materiali: upload, parsing, chunking, metadata	Fabbrica (batch)
Tutor	P0	Spiegazioni guidate, TUTOR MODE, sequenza didattica	Bibliotecario 80% + Ass. Senior 15%
Socrates AI	P0	Dialogo maieutico, domande progressive, SOCRATES MODE	Ass. Senior 40% + Professore 30%
Test con Quiz	P0	Quiz procedurali, micro-verifiche, feedback distrattore	Bibliotecario 90% (pre-generati)
Simulazione Esame	P1	Prove realistiche, previsione voto, feedback dettagliato	Professore 60% + Ass. Senior 30%
Life Planner	P2	Gap analysis + piano settimanale (attivabile a metà pilot)	Ass. Senior
Nexus Graph	P2	Visualizzazione Knowledge Graph (frontend puro)	Nessun LLM

2. COWORK: L'ACCELERATORE DI SVILUPPO

Anthropic ha rilasciato Claude Cowork il 12 gennaio 2026 (macOS) e il 10 febbraio 2026 (Windows). È un agente AI desktop basato su Claude Opus 4.6 che automatizza task multi-step su file locali, con plugin specializzati e integrazione MCP. Per il pilot QWIKKEN, Cowork rappresenta un moltiplicatore di produttività che riduce il fabbisogno di FTE tradizionali.

2.1 Dove Cowork Accelera lo Sviluppo

Area	Task Tradizionale	Ore Stimate (Senza)	Con Cowork + Claude Code	Ore Stimate (Con)	Riduzione
Prompt Engineering	Scrivere, testare, iterare 36.000 prompt per la Fabbrica	120h	Template parametrici + batch generation automatizzata	30h	75%
QA Artefatti	Revisione manuale di 3.600 artefatti (10% del totale)	80h	Cowork analizza consistenza, formato, copertura syllabus su file locali	20h	75%
API Backend	Scrivere routing engine 4-strati, endpoint REST, integrazione DB	200h	Claude Code genera boilerplate + Cowork crea test e documentazione	80h	60%
Frontend Integration	Collegare frontend QWIKKEN BASE ai nuovi endpoint	120h	Claude Code + chrome agent per test E2E automatizzati	50h	58%
Schema DB	Progettare tabelle artefatti, indici, migration scripts	40h	Cowork genera DDL, migration, seed data da spec	10h	75%
Documentazione	Scrivere doc tecnica, API reference, guide onboarding	60h	Cowork genera da codebase + commenti inline	12h	80%
Testing / Debug	Unit test, integration test, load test	100h	Claude Code genera test suite; Cowork organizza e reporta risultati	35h	65%
Monitoraggio	Setup dashboards, alerting, metriche custom	30h	Cowork configura Cloud Monitoring da template	8h	73%
TOTALE		750h		245h	67%

Cowork + Claude Code riducono lo sforzo di sviluppo complessivo del 67%: da 750 ore-uomo a 245 ore-uomo. Questo significa che il team di 5 persone può completare il pilot in 16 settimane lavorando al 60-70% della capacità, oppure — più realisticamente — il pilot può essere realizzato con 3-4 persone al posto di 5-6.

2.2 Stack AI Tooling del Team

Tool	Licenza	Costo Mensile	Utenti	Costo 4 Mesi	Utilizzo
Claude Max (con Cowork)	$200/mese	€185/utente	3 (CTO, ML, Dev)	€2.220	Cowork per automazione, Opus 4.6 per reasoning
Claude Pro	$20/mese	€18.50/utente	1 (Product)	€74	Task leggeri, documentazione, comunicazione
Claude Code	Incluso in Max	€0	2 (CTO, Dev)	€0	Coding agentico, refactoring, test generation
Google AI Studio	Free tier	€0	Tutto il team	€0	Test prompt, prototipazione rapida
SUBTOTALE AI TOOLING				€2.294	Per tutto il pilot

3. BUDGET COMPLETO: Team + Infrastruttura + LLM

3.1 Costo Team (4 Mesi di Pilot)

Le RAL sono quelle previste nel Business Plan (MasterDoc, Cap. 41). Il costo azienda include contributi e TFR stimati al 35% della RAL. L'allocazione indica la percentuale del tempo dedicata al pilot.

Ruolo	RAL	Costo Azienda/Anno	Allocazione Pilot	FTE equiv.	Costo 4 Mesi
CEO / Founder	€60.000	€81.000	30% (sales, partner, CEO)	0.30	€8.100
CTO / Tech Lead	€55.000	€74.250	90% (architettura, Fabbrica, routing)	0.90	€22.275
Full Stack Developer	€45.000	€60.750	85% (API, frontend, integrazione)	0.85	€17.213
ML / AI Engineer	€50.000	€67.500	80% (prompt eng., QA artefatti, tuning)	0.80	€18.000
Product / Customer Success	€40.000	€54.000	50% (onboarding, support, metriche)	0.50	€9.000
TOTALE TEAM	€250.000	€337.500	Media: 67%	3.35 FTE	€74.588

Nota sull'impatto di Cowork: senza Cowork e Claude Code, il pilot richiederebbe ~750h di sviluppo, equivalenti a ~4.7 FTE per 4 mesi. Con l'AI tooling, l'effort scende a ~245h = 1.5 FTE per il coding puro + le altre attività (sales, onboarding, QA). Il team attuale di 5 persone al 67% medio (3.35 FTE) è **sufficiente senza assunzioni aggiuntive**.

3.2 Costi Infrastruttura Cloud (4 Mesi)

Voce	Costo Mensile	Costo 4 Mesi	Note
Cloud Run (backend API)	€45	€180	Auto-scaling 1-10 istanze, 1 vCPU
Cloud SQL (PostgreSQL + pgvector)	€60	€240	db-g1-small, backup giornaliero
Memorystore Redis	€30	€120	Cache sessioni + context prefix
Cloud Storage	€5	€20	Documenti corso + artefatti JSON
Firebase Auth	€0	€0	Free tier (< 50K utenti)
Monitoring + Logging	€10	€40	Cloud Monitoring con alerting
SUBTOTALE INFRA	€150	€600	

3.3 Costi LLM (Inference Runtime + Fabbrica)

Strato	% Traffico	Costo/Mese	Costo 4 Mesi	Note
Fabbrica (batch, una tantum)	—	€891 (1x)	€891	36.000 artefatti, Gemini 2.5 Pro Batch API
Bibliotecario (2.0 Flash)	77%	€55	€220	Con context caching -90%
Assistente Senior (3 Flash)	15%	€120	€480	Escalation media
Professore (3 Pro)	8%	€180	€720	Con context caching -70% input
SUBTOTALE LLM		€355/mese	€2.311	

3.4 Budget Totale del Pilot

Categoria	Costo 4 Mesi	% del Totale	Note
Team (5 persone, 3.35 FTE)	€74.588	87.2%	RAL da Business Plan + contributi 35%
Infrastruttura Cloud	€600	0.7%	Google Cloud (Run, SQL, Redis, Storage)
LLM Runtime (Inference)	€1.420	1.7%	Bibliotecario + Ass. Senior + Professore
LLM Fabbrica (una tantum)	€891	1.0%	Pre-generazione 36.000 artefatti
AI Tooling Team (Cowork + Claude)	€2.294	2.7%	3 licenze Max + 1 Pro
Contingency 10%	€5.720	6.7%	Buffer per imprevisti
BUDGET TOTALE PILOT	€85.513	100%	

> *Il costo reale del pilot è €85.500, di cui l'87% è il team. I costi vivi (infrastruttura + LLM + tooling) sono solo €5.205 — il 6% del totale. Il team è già in organico, quindi il costo incrementale reale per lanciare il pilot è di ~€5.200 + il costo opportunità di 3.35 FTE per 4 mesi.*

3.5 Impatto Economico di Cowork sullo Sviluppo

Cosa succederebbe se non usassimo Cowork e Claude Code? Il team attuale non sarebbe sufficiente: servirebbe almeno 1 developer aggiuntivo per rispettare la timeline di 16 settimane.

Voce	Senza AI Tooling	Con Cowork + Claude Code	Delta
Ore di sviluppo richieste	750h	245h	-505h (67%)
FTE necessari (4 mesi)	4.7 FTE dev + 1.5 FTE altro = 6.2 FTE	1.5 FTE dev + 1.85 FTE altro = 3.35 FTE	-2.85 FTE
Assunzioni aggiuntive necessarie	1 Full Stack Developer (€45K RAL)	Nessuna	Risparmio €20.250 (4 mesi)
Costo AI Tooling	€0	€2.294	+€2.294
Costo team totale (4 mesi)	€94.838 (6 persone)	€74.588 (5 persone)	-€20.250
Timeline stimata	18-20 settimane	16 settimane	-2-4 settimane
RISPARMIO NETTO			€17.956 + 2-4 settimane

Cowork si ripaga 8x: €2.294 investiti in licenze Claude generano €17.956 di risparmio netto (mancata assunzione) + 2-4 settimane di accelerazione. E questo senza contare la qualità superiore del codice generato (test automatici, documentazione inline, consistenza architetturale).

4. TIMELINE: 16 Settimane

Sett.	Fase	Deliverable	Tool AI Utilizzati	Owner
1-2	Setup + Onboarding	Setup Cloud, schema DB, onboarding 6 syllabus, validazione docente	Claude Code (infra-as-code), Cowork (DDL, migration)	CTO + ML
3-4	La Fabbrica	Batch: 36.000 artefatti. QA 10%. Docente valida 1 corso	Claude Code (prompt template), Cowork (QA batch su file)	ML Engineer
5-6	Bibliotecario + Routing	API 4-strati, intent classification, test 50 beta	Claude Code (endpoint + test suite), Cowork (doc API)	Dev + CTO
7-8	Assistente Senior + Socrates	Escalation Gemini 3 Flash, SOCRATES MODE, micro-verifiche	Claude Code (routing logic), Cowork (test E2E)	ML + Dev
9-10	Soft Launch 500	Monitoring real-time, dashboard, iterazione prompt	Cowork (dashboard config, report automatici)	Tutto il team
11-12	ARENA + Scale 1.200	Simulazione Esame attiva, Learning Loop operativo	Claude Code (scoring engine), Cowork (analisi feedback)	ML + Product
13-14	Full Scale 2.000	Tutti gli studenti, ottimizzazione caching su dati reali	Cowork (analisi costi, ottimizzazione query)	CTO + Product
15-16	Valutazione + Report	Metriche finali, report ateneo, proposta commerciale	Cowork (genera report, slide, analisi dati)	CEO + Product

5. METRICHE DI SUCCESSO

5.1 Adozione

KPI	Target Min.	Target Ottimo	Misurazione
DAU (Daily Active Users)	30% (600)	50% (1.000)	Login + 1+ query/giorno
WAU (Weekly Active Users)	60% (1.200)	80% (1.600)	Login + 3+ query/settimana
Retention Week 4	>60%	>80%	% attivi sett. 1 ancora attivi sett. 4
Retention Week 12	>40%	>65%	% ancora attivi a 3 mesi

5.2 Efficacia Didattica

KPI	Target Min.	Target Ottimo	Misurazione
Mastery media fine pilot	>60% UC studiate	>75%	Score da micro-verifiche
Miglioramento Quiz	+15% dal DCA	+30%	Delta quiz iniziale vs fine pilot
NPS studente	>30	>50	Survey sett. 8 e sett. 16
Soddisfazione risposta	>3.5/5	>4.2/5	Rating inline dopo interazioni
Previsione voto vs voto reale	Correlazione >0.6	>0.8	Se disponibile voto esame

5.3 Tecniche ed Economiche

KPI	Target	Allarme Se
Latenza P95 Bibliotecario	<800ms	>2s
Latenza P95 Assistente Senior	<2s	>4s
Latenza P95 Professore	<4s	>8s
Escalation al Professore	<10% dopo sett. 8	>20%
Context Cache hit rate	>90%	<70%
COGS reale / studente / mese	<€5.50	>€8.00
Uptime	>99.5%	<99%

6. FLUSSI UTENTE NEL PILOT

6.1 Flusso TUTOR (Studio Guidato)

Step	Studente	Sistema	Strato OIOD
1	Seleziona corso e argomento	Carica contesto UC dalla Fabbrica	— (DOC)
2	“Voglio studiare il BEP”	Seleziona artefatto livello L2 (dal DCA)	Bibliotecario
3	“Non ho capito”	Serve analogia alternativa #2	Bibliotecario
4	“E con i numeri?”	Serve esempio numerico pre-generato	Bibliotecario
5	Completa lettura	Propone micro-verifica quiz	Bibliotecario
6a	Risponde correttamente	Mastery +, prossimo concetto	Bibliotecario
6b	Risponde sbagliato	Feedback per distrattore, riformulazione	Bibliotecario
7	Domanda inaspettata	FAQ non coperta → escalation	Assistente Senior
8	Collegamento cross-concept	Non pre-generato → escalation	Professore

6.2 Flusso SOCRATES (Maieutica)

Step	Studente	Sistema	Strato
1	“Sfidami sul BEP”	Carica albero socratico pre-generato (3-4 livelli)	Bibliotecario
2-3	Risponde alle domande guida	Branching dell’albero, hint calibrati	Bibliotecario
4	Albero esaurito	Passa a reasoning adattivo	Assistente Senior
5	Ragionamento libero	Dialogo genuino su implicazioni	Professore
6	Sessione conclusa	Se rating alto, risposta promossa ad artefatto	Learning Loop

6.3 Flusso ARENA (Simulazione Esame)

Step	Studente	Sistema	Strato
1	Seleziona corso per simulazione	Genera prova calibrata (30% facili, 50% medie, 20% difficili)	Professore
2	Risponde alle domande	Valutazione automatica (chiuse) + LLM (aperte)	Bibliotecario / Ass. Senior
3	Completa simulazione	Score, previsione voto, gap analysis	Assistente Senior
4	Chiede feedback su un errore	Spiegazione dettagliata con collegamento ai concetti	Professore

7. RISCHI E MITIGAZIONI

Rischio	Prob.	Impatto	Mitigazione
Ateneo non fornisce materiali in tempo	Alta	Bloccante	Pre-accordo. Solo corsi con materiali già digitali
Adozione studenti bassa (<20% DAU)	Media	Alto	Onboarding guidato + incentivo: accesso esclusivo simulazione esame
Qualità artefatti insufficiente	Media	Alto	QA 10% + validazione docente 1 corso + feedback studente dal giorno 1
Costi LLM sopra stime	Bassa	Medio	Budget ha contingency 10%. Monitoring billing giornaliero
Esperienza “robotiche”	Media	Alto	Bibliotecario riformula. Varianti multiple. Randomizzazione
Docenti non validano	Alta	Medio	Validazione incentivata ma non bloccante. Feedback studente come proxy
Rate limit Vertex AI	Molto Bassa	Alto	80 QPM picco vs 1.000 RPM limit = margine >10x
Limiti Cowork (usage cap)	Media	Basso	Rotazione sessioni tra team. Claude Code per coding puro. Budget per upgrade se necessario

8. DELIVERABLE PER L'ATENEO

Deliverable	Contenuto	Consegna
Report Pilot	Metriche adozione, efficacia didattica, costi reali vs stime, lesson learned	Settimana 16
Dashboard Live	KPI in tempo reale durante il pilot	Dalla settimana 9
Analisi Economica	COGS reale, proiezione 10K-50K studenti, TCO per ateneo	Settimana 16
Proposta Commerciale	Pricing definitivo, SLA, roadmap funzionalità	Settimana 16-17
Knowledge Base	36.000 artefatti validati per 6 corsi	Fine pilot
Roadmap Moduli	Life Planner, Nexus Graph, Thesis Forge, Linguistic Gym	Settimana 16

9. SCENARIO COMMERCIALE POST-PILOT

Fase	Timeline	Studenti	Revenue	COGS/Stud.	Margine
Pilot (questo doc)	Mesi 1-4	2.000	€0 (gratuito)	N/A	Investimento
Early Adoption	Mesi 5-8	5.000	€500K	€20.62	79%
Primo contratto	Mesi 9-12	15.000	€1.5M	€16.00	84%
Scale Anno 2	Mesi 13-24	45.000	€4.5M	€12.00	88%

Il pilot è un investimento di €85.500 fully loaded (di cui €74.600 è team già in organico e €5.200 sono costi incrementali). Se converte in un contratto da 15.000 studenti a €99.99, il revenue annuo è €1.5M con margine lordo 84%. **ROI del pilot: 17x sui costi vivi, 18x sul P&L se si considerano i soli costi incrementali.**

10. NEXT STEPS

#	Azione	Owner	Deadline
1	Identificare ateneo partner (primo contatto o referral)	CEO	Settimana 1
2	Ottenere 6 syllabus + materiali digitali (PDF/DOCX)	CEO + Ateneo	Settimana 2-3
3	Setup Google Cloud + attivare Vertex AI	CTO	Settimana 1
4	Configurare Cowork con plugin dev + folder di progetto	CTO	Settimana 1
5	Ingestione materiali (SENSORIUM) + mapping UC (SYNAPSE)	ML Engineer	Settimana 2-3
6	Lancio Fabbrica: batch 36.000 artefatti	ML + CTO	Settimana 3-4
7	QA artefatti + validazione docente	ML + Docente	Settimana 4-5
8	API routing 4-strati + integrazione QWIKKEN BASE	Dev + CTO	Settimana 3-6
9	Beta test 50 utenti	Team	Settimana 6-7
10	Soft launch 500 → 1.200 → 2.000	Product + CTO	Settimana 9-13

Il pilot può partire oggi. Il frontend esiste (QWIKKEN BASE Rev 5.0). L’architettura OIOD è progettata. Cowork accelera lo sviluppo del 67%. I prezzi Google sono confermati. Il costo incrementale è ~€5.200. L’unica variabile è l’ateneo partner.