QWIKKEN — Offline Intelligence, Online Delivery

Architettura di Pre-Generazione Artefatti per l'Ottimizzazione dei Costi AI

Rev. 2.0 — Prezzi Reali Febbraio 2026
Documento Integrativo al SOCRATES 2.0
18 Febbraio 2026
Synthetic Data S.r.l. — Confidenziale

CHANGELOG

Rev.	Data	Modifiche
1.0	12 Feb 2026	Documento originale. Architettura 3 strati (Fabbrica/Bibliotecario/Professore). Stime costi basate su prezzi stimati.
2.0	18 Feb 2026	Aggiornamento con prezzi reali Google feb 2026. Introduzione 4° strato (Gemini 3 Flash). Scenario Pilot Zero-Cost per CEPO Multiversity. Ricalcolo COGS completo.

1. EXECUTIVE SUMMARY

La Rev. 2.0 di questo documento aggiorna l'architettura Offline Intelligence, Online Delivery con i prezzi reali dei modelli Google a febbraio 2026, introduce un quarto strato architetturale (Gemini 3 Flash) e include uno scenario di Pilot a costo quasi-zero per la demo CEPO Multiversity.

Risultato chiave: il COGS per studente scende da €40 (architettura attuale) a **€15-19 con la nuova architettura a 4 strati**, grazie alla combinazione di pre-generazione offline, Batch API al 50% di sconto, Context Caching al 90% di riduzione, e Gemini 3 Flash come strato intermedio.

Metrica	Arch. Attuale	Rev 1.0 (3 strati)	Rev 2.0 (4 strati)	Delta vs Attuale
COGS Inferenza / studente / anno	€26.00	€12.00	€6.50-9.00	−65-75%

Metrica	Arch. Attuale	Rev 1.0 (3 strati)	Rev 2.0 (4 strati)	Delta vs Attuale
COGS Totale / studente / anno	€40.00	€27.10	€15.50-19.00	−52-61%
Margine Lordo	60%	73%	81-85%	+21-25pp
% traffico su Pro (runtime)	15%	15-20%	3-5%	−10-12pp
% traffico su Flash 3 (runtime)	—	—	10-15%	Nuovo strato
% servito da artefatti pre-generati	0%	70-80%	75-85%	+75-85pp
Costo Fabbrica (Batch API)	—	€0.10	€0.04-0.06	−40-60%

2. PREZZI REALI: Listino Google Febbraio 2026

2.1 Listino Modelli Google (Febbraio 2026)

Modello	Input / 1M tok	Output / 1M tok	Context Window	Ruolo in QWIKKEN
Gemini 2.0 Flash-Lite	$0.10	$0.40	1M tokens	Bibliotecario (opzione economy)
Gemini 2.0 Flash	$0.10	$0.40	1M tokens	Bibliotecario (default)
Gemini 2.5 Flash	$0.30	$2.50	1M tokens	Bibliotecario (reasoning ibrido)
Gemini 3 Flash Preview	$0.50	$3.00	1M tokens	**Assistente Senior (NUOVO)**
Gemini 2.5 Pro	$1.25	$10.00	1M tokens	Professore (standard)
Gemini 3 Pro Preview	$2.00	$12.00	1M tokens	Professore (top tier)

2.2 Sconti Strutturali Confermati

Sconto	Meccanismo	Riduzione	Applicazione in QWIKKEN
Batch API	Processamento asincrono	-50% su tutti i modelli paid	LA FABBRICA: tutta la pre-generazione
Context Caching	Cache del prefix di contesto	-90% sui token di input cachati	BIBLIOTECARIO + PROFESSORE
Context Caching Storage	Mantenimento cache in memoria	$1-4.50 / 1M tok / ora	Costo marginale

Sconto	Meccanismo	Riduzione	Applicazione in QWIKKEN
Free Tier (AI Studio)	Accesso gratuito con rate limit	100% gratis (15 RPM, 1K RPD)	PILOT: primi 200-500 studenti

> **Insight critico:** il Batch API al 50% di sconto rende la Fabbrica 2x più economica di quanto stimato nella Rev. 1.0. Con Gemini 2.5 Pro in batch, il costo di pre-generazione scende a 0.625/5.00 per milione di token — circa €60 per un intero corso di 200 Unità Concettuali.

3. ARCHITETTURA REV 2.0: Quattro Strati

La principale innovazione della Rev. 2.0 è l'introduzione di un quarto strato: l'Assistente Senior, basato su Gemini 3 Flash. Questo modello, rilasciato a fine 2025, offre intelligenza di livello frontier a un prezzo 4x inferiore al Pro, creando un buffer naturale tra il Bibliotecario economico e il Professore costoso.

Strato	Nome	Modello AI	Costo In/1M	Costo Out/1M	Quando Opera	% Interazioni
1	**LA FABBRICA**	Gemini 2.5 Pro (Batch)	$0.625	$5.00	Offline (batch)	0% (produce artefatti)
2	**IL BIBLIOTECARIO**	Gemini 2.0 Flash	$0.10	$0.40	Runtime (sempre)	75-80%
3	**L'ASSISTENTE SENIOR**	Gemini 3 Flash	$0.50	$3.00	Runtime (escalation media)	12-18%
4	**IL PROFESSORE**	Gemini 3 Pro	$2.00	$12.00	Runtime (escalation alta)	3-5%

3.1 Il Nuovo Strato: L'Assistente Senior (Gemini 3 Flash)

Il gap tra Gemini 2.0 Flash (0.10/0.40) e Gemini 3 Pro (2.00/12.00) è di 20x sull'input e 30x sull'output. Nella Rev. 1.0 qualsiasi escalation dal Bibliotecario saltava direttamente al Professore, bruciando budget. Gemini 3 Flash si inserisce nel mezzo con un rapporto costo/intelligenza eccezionale.

3.1.1 Cosa Gestisce l'Assistente Senior

Caso	Descrizione	Perché Non Basta il Bibliotecario	Perché Non Serve il Professore
Domanda fuori FAQ (semplice)	Domanda non coperta dalle FAQ ma	Richiede generazione, non solo lookup	Non serve reasoning multi-step

Caso	Descrizione	Perché Non Basta il Bibliotecario	Perché Non Serve il Professore
	risolvibile con contesto corso		
Riformulazione avanzata	Studente non capisce dopo 2 tentativi	Richiede creatività nella spiegazione	Il concetto è noto, serve solo altra angolazione
SOCRATES livello 2-3	Albero socratico pre-generato esaurito	Serve reasoning adattivo	Non serve cross-concept analysis
Correzione esercizi	Feedback dettagliato sul procedimento	Serve analisi del ragionamento	Pattern riconoscibile
Collegamento prerequisiti	Come un concetto si collega a uno precedente	Serve navigazione del Knowledge Graph	I concetti sono noti, serve solo la connessione

3.1.2 Impatto Economico del Quarto Strato

Metrica	Rev 1.0 (al Professore)	Rev 2.0 (all'Assistente Senior)	Risparmio
Costo medio per interazione (1K in + 500 out)	$0.0070	$0.0020	71%
Interazioni / studente / anno (stimate)	480	480	—
Costo annuo per 12% traffico	€3.02	€0.86	€2.16/studente
Su 70.000 studenti (Anno 5)	€211K	€60K	€151K risparmiati

3.2 La Fabbrica: Ricalcolo con Batch API

Artefatto	Volume (per 200 UC)	Token Medi	Modello Batch	Costo Unitario	Costo Totale
Spiegazioni Multi-Livello	800	~1.500	2.5 Pro Batch	$0.0045	€33.00
Analogie Alternative	600	~800	2.5 Pro Batch	$0.0025	€13.80
Esempi Numerici	600	~1.200	3 Flash Batch	$0.0012	€6.60
Micro-Verifiche	800	~600	3 Flash Batch	$0.0006	€4.40
FAQ Strutturate	2.400	~1.000	2.5 Pro Batch	$0.0035	€77.00

Artefatto	Volume (per 200 UC)	Token Medi	Modello Batch	Costo Unitario	Costo Totale
Alberi Socratici	200	~2.000	2.5 Pro Batch	$0.0065	€12.00
Contestualizzazioni + Schede	600	~500	2.0 Flash	$0.0003	€1.60
TOTALE PER CORSO	**~6.000 artefatti**				**€148.40**

Ammortamento: 2.000 studenti = €0.07/studente. 200 studenti = €0.74/studente. Irrisorio.

> Con il Batch API, la Fabbrica costa €148 per un corso intero di 200 UC. Rispetto al costo annuo di inferenza runtime di €26/studente, la pre-generazione si ripaga con il primo studente che la utilizza.

3.3 Il Bibliotecario: Costi con Context Caching

Scenario	Token Input	Token Output	Costo / Query	Costo / Studente / Anno (3.200 query)
Senza caching	20K	500	$0.0022	€6.40
Con Context Caching (90% hit)	2K + 18K cached	500	$0.00056	€1.63
Con caching + semantic cache (40% hit)	Effettivo ~1.2K	300	$0.00028	€0.81

Risparmio del Bibliotecario: da €6.40/studente/anno a **€0.81/studente/anno** con doppio livello di caching. Riduzione dell'87%.

3.4 Copertura Aggiornata per Modalità

Modalità SOCRATES 2.0	Bibliotecario	Assistente Senior	Professore	Nota
TUTOR MODE (55% traffico)	90-95%	3-7%	1-3%	Quasi tutto pre-tracciato
RESPONDER MODE (30% traffico)	55-65%	25-30%	5-10%	FAQ + Assistente per imprevisti
SOCRATES MODE (15% traffico)	25-35%	30-40%	25-35%	Alberi pre-generati + reasoning
MEDIA PONDERATA	**~77%**	**~15%**	**~8%**	Mix distribuzione reale

4. RICALCOLO COGS: Unit Economics Febbraio 2026

4.1 Profilo Studente Tipo (invariato)

Parametro	Valore	Note
Periodo attività	8 mesi/anno	Sessioni d’esame + lezioni
Esami/anno	6	Media studente telematica
Interazioni/giorno (attivo)	20 turni	Mix TUTOR/RESPONDER/SOCRATES
Giorni attivi/anno	~160	8 mesi × 20 giorni
Interazioni totali/anno	~3.200	160 × 20

4.2 Costo Inferenza per Strato

Strato	% Interaz.	Query/Anno	Tok In (con cache)	Tok Out	Costo/Query	Costo/Stude
Bibliotecario (2.0 Flash + cache)	77%	2.464	2K eff.	400	$0.00036	€0.81
Assistente Senior (3 Flash)	15%	480	5K	600	$0.0043	€1.88
Professore (3 Pro + cache)	8%	256	5K eff.	800	$0.0196	€4.56
TOTALE INFERENZA	**100%**	**3.200**				**€7.25**

> Il costo di inferenza runtime scende da €26/studente (architettura attuale) a €7.25/studente — una riduzione del 72%.

4.3 COGS Totale per Studente

Voce di Costo	Arch. Attuale	Rev 1.0	Rev 2.0	Note Rev 2.0
Ingestione (SENSORIUM)	€6.00	€6.00	€5.00	Flash-Lite per parsing, Batch per embedding
Memoria (AlloyDB + pgvector)	€6.00	€7.00	€6.50	Incluso storage artefatti
Pre-generazione (Fabbrica)	—	€0.10	€0.07	Batch API -50%
Inferenza Runtime	€26.00	€12.00	€7.25	4 strati + caching
Infrastruttura (Cloud Run, Redis)	€2.00	€2.00	€1.80	Ottimizzazione scaling
TOTALE COGS	**€40.00**	**€27.10**	**€20.62**	

Voce di Costo	Arch. Attuale	Rev 1.0	Rev 2.0	Note Rev 2.0
PREZZO	**€99.99**	**€99.99**	**€99.99**	
MARGINE LORDO	**€59.99 (60%)**	**€72.89 (73%)**	**€79.37 (79%)**	**+19pp vs attuale**

5. IL CICLO VIRTUOSO: Learning Loop Potenziato

Proiezione di Convergenza (Rev 2.0)

Mese	% Bibliotecario	% Assistente Senior	% Professore	COGS Inferenza/Studente
Mese 1 (lancio)	77%	15%	8%	€7.25
Mese 3	82%	12%	6%	€5.80
Mese 6	86%	10%	4%	€4.50
Mese 12 (regime)	89%	8%	3%	€3.60
Mese 24 (maturo)	92%	6%	2%	€2.80

A regime (mese 12+): il COGS inferenza scende a ~€3.60/studente/anno. Il COGS totale converge verso **€14-16/studente**, portando il margine lordo all'84-86%.

6. SENSITIVITY ANALYSIS

6.1 Scenari di Costo LLM

Scenario	COGS	Margine Lordo	Nota
Base Rev 2.0 (prezzi attuali)	€20.62	79%	Conservativo, anno 1
Costi LLM +50%	€24.25	76%	Ancora eccellente
Costi LLM +100%	€27.87	72%	Pari alla Rev 1.0 base
Costi LLM +200%	€35.12	65%	Ancora sopra soglia 60%
Costi LLM -30% (trend 2027)	€16.43	84%	Probabile per calo naturale
Con Gemma fine-tuned (Anno 3)	€12.00	88%	Bibliotecario self-hosted
Regime (Learning Loop mese 12)	€16.20	84%	Solo ottimizzazione interna

Resilienza: anche con un raddoppio dei costi LLM, l'architettura Rev 2.0 mantiene un margine lordo del 72% — superiore al margine base dell'architettura attuale (60%).

6.2 Proiezione 5 Anni

Anno	Studenti	COGS Unit.	COGS Tot.	Revenue	Margine Lordo	Margine %
2026 (Pilot)	2.000	€20.62	€41K	€200K	€159K	79%
2027	15.000	€16.00	€240K	€1.5M	€1.26M	84%
2028	45.000	€12.00	€540K	€4.5M	€3.96M	88%
2029	80.000	€9.00	€720K	€8.0M	€7.28M	91%
2030	100.000	€7.00	€700K	€10.0M	€9.30M	93%

7. SCENARIO PILOT: Costo Quasi-Zero per CEPO Multiversity

Questo scenario è progettato specificamente per la demo CEPO: dimostrare che il pilot iniziale (200-500 studenti, 1-2 corsi) può partire con un costo infrastrutturale minimo.

7.1 Risorse Gratuite Disponibili

Risorsa	Free Tier	Limite	Sufficiente per Pilot?
Google AI Studio (2.0 Flash)	Gratuito	15 RPM, 1.000 RPD, 250K TPM	Sì per 200 studenti a basso carico
Google AI Studio (3 Flash Preview)	Gratuito	5 RPM, 100 RPD	Sufficiente per escalation limitate
Firebase Auth	Gratuito fino a 50K utenti	50.000 auth/mese	Più che sufficiente
Cloud SQL (trial)	€300 credit Google Cloud	90 giorni	Copre il pilot completo
Cloud Run	€300 credit + free tier	2M richieste/mese gratis	Più che sufficiente

7.2 Architettura Pilot (200 Studenti, 1 Corso)

Componente	Servizio	Costo Mensile	Note
LLM Inference (Bibliotecario)	AI Studio Free (2.0 Flash)	€0	1.000 RPD coprono ~200 studenti × 5 query/giorno

Componente	Servizio	Costo Mensile	Note
LLM Inference (Assistente Senior)	AI Studio Free (3 Flash)	€0	100 RPD per escalation (~5% traffico)
LLM Inference (Professore)	Vertex AI Paid (overflow)	~€15-30/mese	Solo per escalation che superano free tier
Pre-generazione (Fabbrica)	Batch API (una tantum)	~€75	1 corso, 200 UC, costo una tantum
Database	Google Cloud Trial Credit	€0 (credit)	Cloud SQL micro instance
Backend (Cloud Run)	Free tier + credit	€0	Ampiamente nel free tier
TOTALE MENSILE PILOT		**€15-30/mese**	**Escluso costo una tantum Fabbrica**

> **Il pilot con 200 studenti costa circa €75 una tantum (Fabbrica) + €15-30/mese (overflow Professore). Costo totale per 3 mesi di pilot: €120-165. Investimento sotto i €200 per validare la tecnologia con studenti reali.**

7.3 Limiti del Pilot e Scalabilità

Limite Free Tier	Impatto sul Pilot	Soluzione per Scale
1.000 RPD su Flash	~5 query/studente/giorno con 200 studenti	Paid Tier sopra i 500 studenti
100 RPD su 3 Flash/Pro	~0.5 escalation/studente/giorno	Sufficiente per pilot; Paid per produzione
Rate limit 15 RPM su Flash	Max 15 studenti simultanei	OK per pilot; 1000+ RPM in Paid
No Batch API su Free Tier	Fabbrica richiede Paid	Costo una tantum irrisorio (€75/corso)
No SLA su Free Tier	Nessuna garanzia uptime	Accettabile per pilot

8. CONFRONTO COMPLETO: Rev 1.0 vs Rev 2.0

Dimensione	Rev 1.0 (Feb 12)	Rev 2.0 (Feb 18)	Miglioramento
Strati architetturali	3 (Fabbrica/Bibliotecario/Professore)	4 (+Assistente Senior)	Buffer costi tra Flash e Pro
Prezzi LLM	Stimati	Reali (confermati Google feb 2026)	Accuratezza +100%
Batch API	Non considerato	Integrato (-50% Fabbrica)	Costo Fabbrica dimezzato

Dimensione	Rev 1.0 (Feb 12)	Rev 2.0 (Feb 18)	Miglioramento
Context Caching	Menzionato ma non calcolato	Calcolato (-87% costo Bibliotecario)	Da €6.40 a €0.81/studente
COGS Totale / studente	€27.10	€20.62	-24%
Margine Lordo Anno 1	73%	79%	+6pp
Margine a Regime	78%	84-86%	+6-8pp
Scenario Pilot	Non incluso	€120-165 per 3 mesi / 200 studenti	Pitch-ready per CEPO
Resilienza (+100% costi LLM)	61%	72%	+11pp

9. PIANO DI IMPLEMENTAZIONE AGGIORNATO

Fase	Timeline	Deliverable	Dipendenze	Costo
0. Pilot CEPO	Sett. 1-4	Demo funzionante con 1 corso, 200 studenti, free tier	Nessuna	€75 (Fabbrica)
1. Schema Artefatti	Sett. 1-2	Schema JSON per ogni tipo + tabelle DB	Nessuna	€0
2. Pipeline Fabbrica	Sett. 3-5	Cloud Run Job con Batch API	Fase 1	€0 (dev)
3. API Bibliotecario + Routing	Sett. 5-8	Endpoint con 4-tier routing	Fase 1	€0 (dev)
4. Integrazione SOCRATES 2.0	Sett. 7-10	Merge con architettura trimodale	Fase 3	€0 (dev)
5. Learning Loop	Sett. 9-12	Pipeline escalation + promozione artefatti	Fasi 2-4	€0 (dev)
6. Dashboard Docente	Sett. 10-14	UI validazione artefatti	Fase 2	€0 (dev)
7. A/B Test Pilot	Sett. 13-16	200 studenti: old vs new arch	Fasi 1-5	~€200
8. Produzione	Sett. 17+	Deploy progressivo	Fase 7	Paid tier Google

Novità Rev 2.0: la Fase 0 (Pilot CEPO) può partire immediatamente in parallelo con lo sviluppo, utilizzando il free tier di Google AI Studio e una Fabbrica semplificata.

10. CONCLUSIONI

Dimensione	Senza OIOD	Rev 1.0	Rev 2.0
Struttura costi	100% variabile	~80% fisso + 20% var.	~85% fisso + 15% var.
Esposizione a var. prezzi LLM	100% traffico esposto	20% esposto	15% esposto
COGS / studente	€40.00	€27.10	€20.62
Margine Lordo Anno 1	60%	73%	79%
Margine a Regime	60% (no improvement)	78%	84-86%
Costo per partire (pilot)	Elevato (infra)	Non stimato	€120-165 per 3 mesi
Resilienza (+100% costi LLM)	34% margine	61% margine	72% margine
Qualità risposte	Variabile (runtime)	Alta (pre-validata)	Alta + Assistente Senior per gap

L'architettura a 4 strati non è solo un'ottimizzazione dei costi. Trasforma la struttura economica di QWIKKEN da startup AI con margini incerti a piattaforma SaaS con margini lordi dell'80%+, resiliente alle fluttuazioni di prezzo dei provider, con un costo di ingresso per il pilot sotto i €200. Questo è il numero che conta per il CEPO Multiversity.