HELLFIRECLOUD

Risorsa gratuita · Studio italiano AI & Cloud

Checklist LLM aziendale 2026

Le 12 domande da farsi *prima* di mandare un Large Language Model in produzione.

Per: CTO, Head of Engineering, Tech Lead
Edizione: Maggio 2026

hellfirecloud.it

Francesco Oghabi · Cloud & AI Consultant

Premessa onesta

> Non esiste una risposta universale alla domanda *"quale LLM scelgo per la mia azienda?"*. Esistono però **12 domande** che, se le rispondi con i tuoi numeri reali *prima* di firmare un contratto o di mettere su un'infrastruttura, ti evitano l'80% degli errori che vedo fare ai clienti.

Questa checklist nasce da 7+ progetti reali con Llama, Mistral, Qwen e GPT-4 in aziende italiane tra il 2024 e il 2026. È in ordine di importanza decrescente: la prima domanda da sola può chiudere o sbloccare il 60% delle opzioni.

Come usarla: rispondi alle 12 domande prendendo appunti su un foglio. Quando arrivi in fondo, la scelta del modello/architettura quasi si decide da sola. Se rimangono 2-3 strade plausibili, prenota la call di chiusura in fondo al documento — mezz'ora gratis per validarle insieme.

1 Quali sono i dati che il modello vedrà — e di chi sono?

Se il modello processerà email dei clienti, scansioni di contratti, dati sanitari o codice sorgente proprietario: **cloud LLM API è quasi sempre fuori discussione** per motivi GDPR, NIS2 e contrattuali. Devi andare on-premise (server tuoi o cloud privato europeo) con modello open source self-hosted.

Se invece il modello processa contenuto pubblico (es. riassunti di articoli, generazione marketing copy), hai tutto lo spazio per scegliere.

> **Decisione che sblocca:** apre o chiude almeno il 60% delle opzioni successive.

2 Quanto stai spendendo oggi nel processo che vuoi automatizzare?

Se il processo costa €5.000/mese in stipendi e tempo, e un'API LLM lo risolve a €200/mese, la matematica è ovvia: vai con l'API. Se il processo costa €500.000/mese, ha senso investire 6 mesi di lavoro per fare fine-tuning su un modello open source con un break-even a 12 mesi.

Calcola sempre il **TCO (Total Cost of Ownership) a 24 mesi**, non a 1 mese.

3 Quante richieste al mese, davvero?

Le API LLM hanno prezzi lineari per token. Self-hosting di un modello su GPU ha un costo sostanzialmente fisso (ammortamento + energia). C'è un punto di break-even.

Regola del pollice 2026:

- **Sotto 10 milioni di token/mese** → API cloud (OpenAI, Anthropic, Google) è quasi sempre più conveniente
- **Tra 10 e 100 milioni token/mese** → dipende dal caso, valuta entrambi
- **Oltre 100 milioni di token/mese** → self-hosting su GPU dedicata vince

4 Latenza accettabile?

Chatbot in real time → max 2-3 secondi per risposta completa o devi usare streaming.
Batch processing notturno → puoi tollerare 30+ secondi per richiesta.

La risposta cambia completamente quale modello e quale infrastruttura sono adatti.

5 Sei disposto a investire in MLOps?

Self-hosting un LLM = devi monitorare GPU, gestire deploy, fare load testing, settare alerting. Se non hai un team che può presidiare l'infrastruttura, l'API cloud è la scelta saggia anche se in apparenza più costosa.

Costo nascosto del self-hosting: 0.5-1 FTE (Full-Time Equivalent) di MLOps engineer dedicato.

6 Italiano nativo o no?

Nel 2026, i modelli che gestiscono meglio l'italiano sono:

- **Proprietari:** GPT-4o, Claude 3.5, Gemini 1.5 Pro
- **Open source:** Llama 3.3 70B, Mistral Large 2, Qwen 2.5 72B, Gemma 2 9B
- **Da evitare per italiano:** Phi-3 (eccellente in inglese, debole in italiano)

Se il tuo caso d'uso è italiano-centric, escludi subito i modelli sotto-performanti.

7 Hai bisogno di accedere ai tuoi dati aziendali (RAG)?

Se la risposta del modello deve basarsi sui *tuoi* documenti (manuali interni, knowledge base, contratti), ti serve un sistema **RAG (Retrieval Augmented Generation)**.

Questa è un'infrastruttura a parte: vector database, retriever, re-ranker. Il modello LLM è solo il pezzo finale, e qui pesa relativamente poco la scelta del modello.

8 Hai un dataset di esempi specifici?

Se hai **1000+ esempi** di "input → output desiderato" del tuo caso d'uso specifico, conviene fare **fine-tuning** di un modello open source piccolo.

Risultato: un modello da 7-13B parametri che batte GPT-4 nel *tuo* task specifico, costando 10-100× meno per inferenza.

Se non hai dataset → niente fine-tuning, usa prompt engineering + RAG.

9 Vendor lock-in: quanto è grave per te?

Costruire tutto il prodotto attorno alle API di OpenAI espone a: aumenti di prezzo, cambi di policy, deprecation di modelli, problemi di disponibilità.

Se questo è un rischio strategico per il business, considera uno **strato di astrazione** (LiteLLM, LangChain) che ti permette di switchare provider in 1 giorno, oppure vai direttamente self-hosted.

10 Quale GPU ti serve (se vai self-hosted)?

Linee guida 2026:

- **RTX 4090 / 5090 (24-32GB):** modelli 7-13B quantizzati Q4/Q5. Perfetto per PoC e workload medi
- **NVIDIA A100 80GB:** modelli 70B in Q4 / 30B in Q8. Workload produzione medi
- **2× H100 80GB:** modelli 70B in fp16 / 405B quantizzati. Workload enterprise

Regola d'oro: mai sotto-dimensionare la VRAM. Una GPU con poca memoria che fa swap costa più di una grande.

11 Hai un piano per gestire le allucinazioni?

Ogni LLM allucina. *Sempre.* Domande da fare al team:

- Hai una validation layer che controlla gli output prima che vadano all'utente?
- Hai un meccanismo di feedback per migliorare nel tempo?
- Hai monitoring per individuare drift?

Se le risposte sono "no", il modello **non è pronto per la produzione**, indipendentemente da quanto sia bravo.

12 Hai definito un metro di successo misurabile?

"Vogliamo usare l'AI" non è un obiettivo.

"Vogliamo ridurre del 40% il tempo medio di risposta del customer service entro 6 mesi" lo è.

Senza un metro di successo, qualunque modello sembrerà funzionare per 2 settimane e poi verrà abbandonato.

In sintesi

Se rispondi onestamente a queste 12 domande, la scelta del modello quasi si decide da sola.

Non sono io ad aver "inventato" questa checklist: è la sintesi degli errori (miei e altrui) visti su 7+ progetti LLM reali con aziende italiane tra il 2024 e il 2026.

Hai dubbi sul tuo caso specifico?

Compila mentalmente le 12 domande applicate al tuo progetto.
Poi prenotiamo una call: .
Ti dico subito dove vedo rischi, dove vedo opportunità, e quale architettura consiglierei.

Scrivimi su WhatsApp →

WhatsApp: +39 351 4437048 · Email: info@hellfirecloud.it · hellfirecloud.it

Francesco Oghabi — Cloud & AI Consultant, fondatore di HellfireCloud SRL. Da 8+ anni progetto e metto in produzione architetture cloud, sistemi ML/AI e infrastrutture GPU per aziende italiane. Specializzato in LLM open source self-hosted, sistemi RAG enterprise, MLOps.