# Linee guida per l'IA nella pubblica amministrazione

# Esempio LCOAI

## Strumento C

# Sommario

# 1. Finalità dello strumento

Il presente strumento fornisce un esempio numerico di applicazione della metrica LCOAI – Levelized Cost of Artificial Intelligence (LCOAI), quale strumento di supporto alla:

- programmazione degli acquisti,
- valutazione comparativa tra modelli di deployment (API, cloud, on-premises),
- determinazione della base d'asta,
- analisi di sostenibilità economica nel periodo preso come riferimento temporale nell'esempio che segue (un anno).

La metrica è mutuata dall'approccio illustrato in letteratura e adattata al contesto della Pubblica Amministrazione.

# 2. Formula utilizzata

Per un orizzonte temporale di **1 anno**, la formula è

LCOAI = (CAPEX + OPEX) / Q

Dove:

- CAPEX = costi di investimento iniziale (integrazione, setup, personalizzazione, infrastruttura)
- OPEX = costi operativi ricorrenti annui
- Q = numero di inferenze valide (output produttivi) nel periodo considerato

L'indicatore è espresso in:

€ per singola inferenza

oppure

€ per 1.000 inferenze

# 3. Caso di studio

**Ipotesi di base**

L'esempio applicativo è riferito a un Comune con le seguenti caratteristiche:

- **Popolazione residente:** 120.000 abitanti
- **Servizi digitali attivi:** anagrafe, tributi, edilizia, SUAP, servizi scolastici, segnalazioni
- **Canali di contatto attivi:** sito istituzionale, sportello digitale, app comunale

Si ipotizza l'introduzione di un assistente virtuale IA a supporto delle richieste informative dei cittadini.

Per stimare il volume di utilizzo si procede come segue:

1) Passaggio 1 – Percentuale di cittadini che utilizza il servizio

Ipotesi prudenziale:

- 40% della popolazione utilizza almeno una volta l'anno il canale digitale assistito

$$120.000 \times 40\% = 48.000$$

Orizzonte: 1 anno

2) Passaggio 2 – Numero medio di richieste per utente

Ipotesi media:

- 4 interazioni/anno per utente

$$48.000 \times 4 = 192.000 \text{ conversazioni annue}$$

3) Da conversazioni a inferenze

Una conversazione con un assistente IA non corrisponde a una sola inferenza.

Una conversazione media può includere:

- domanda iniziale
- chiarimento
- risposta di dettaglio
- eventuale riformulazione

Ipotesi realistica:

- 10–12 scambi medi per conversazione
- Ogni scambio genera 1 inferenza

$$192.000 \times 12 = 2.304.000 \text{ inferenze}$$

4) Fattore di crescita e utilizzo sistemico

Nel calcolo LCOAI si considera un valore superiore (8.000.000 inferenze) per tre motivi:

1. Accessi multipli per pratica complessa
2. Integrazione con altri servizi automatici interni
3. Utilizzo anche da parte del personale interno

Ulteriori componenti prevedono:

- chatbot per i cittadini (2.300.000 inferenze stimate);
- supporto agli operatori interni (1.200.000 inferenze stimate);
- gestione automatica di testi e/o atti (1.500.000 inferenze stimate);
- sistemi di classificazione automatica (3.000.000 inferenze stimate);

per un totale stimato di 8.000.000 di inferenze.

**Ulteriori componenti**

| **Fonte** | **Inferenze stimate** |
|---|---|
| **Chatbot cittadini** | 2.300.000 |
| **Supporto operatori interni** | 1.200.000 |
| **Generazione automatica testi/atti** | 1.500.000 |
| **Sistemi di classificazione automatica** | 3.000.000 |
| **Totale stimato** | **8.000.000** |

Vengono confrontati due scenari:

- Scenario A: API esterna (modello SaaS)
- Scenario B: Modello self-hosted su infrastruttura dedicata

**Scenario A (API esterna)**

**CAPEX**

Le voci di costo prevedono:

- Integrazione sistemi, per € 30.000
- Configurazione sicurezza e compliance, per € 10.000
- Test e avviamento, per € 10.000.

In totale, quindi, i CAPEX ammontano a € 50.000

| Voce | Importo (€) |
|---|---|
| Integrazione sistemi | 30.000 |
| Configurazione sicurezza e compliance | 10.000 |
| Test e avviamento | 10.000 |
| **Totale CAPEX** | **50.000** |

**OPEX**

Costo per inferenza (media input/output token): € 0,009 per inferenza

**Calcolo OPEX:**

Il costo per il consumo API è pari a

0,009×8.000.000=72.000€

Il costo per attività di monitoraggio e logging è pari a € 8.000

Il costo del personale per la supervisione è pari a € 10.000.

In totale, quindi, gli OPEX ammontano a € 90.000.

| Voce | Importo (€) |
|---|---|
| Consumo API | 72.000 |
| Monitoraggio e logging | 8.000 |
| Personale supervisione | 10.000 |
| **Totale OPEX** | **90.000** |

**Calcolo LCOAI scenario A**

50.000 + 90.000/8.000.000

LCOAI = 140.000/8.000 = 0,0175 €

Risultato:

- € 0,0175 per inferenza
- € 17,50 per 1.000 inferenze

**Scenario B (Self-hosted)**

**CAPEX**

Le voci di costo prevedono:

- Infrastruttura GPU, per € 120.000
- Setup della data pipeline, per € 30.000
- Fine-tuning del modello, per € 25.000
- Integrazione dei sistemi, per € 25.000.

In totale, quindi, i CAPEX ammontano a € 200.000.

| Voce | Importo (€) |
|---|---|
| Infrastruttura GPU | 120.000 |

| | |
|---|---|
| Setup data pipeline | 30.000 |
| Fine-tuning modello | 25.000 |
| Integrazione sistemi | 25.000 |
| **Totale CAPEX** | **200.000** |

**OPEX**

Costo computazionale medio stimato:

€ 0,0045 per inferenza

Calcolo: 0,0045×8.000.000=36.000€

Il costo per consumo energetico e calcolo è pari a € 36.000

Il costo per la manutenzione dell'infrastruttura è pari a € 20.000

Il costo di ri-addestramento è pari a € 15.000

Il costo del personale tecnico è pari a € 25.000.

In totale, quindi, gli OPEX ammontano a € 96.000.

| **Voce** | **Importo (€)** |
|---|---|
| Consumo energetico e calcolo | 36.000 |
| Manutenzione infrastruttura | 20.000 |
| Ri-addestramento | 15.000 |
| Personale tecnico | 25.000 |
| **Totale OPEX** | **96.000** |

**Calcolo LCOAI scenario B**

200.000 + 96.000/8.000.000

LCOAI = 296.000/8.000 = 0,037 €

Risultato:

- € 0,0175 per inferenza
- € 17,50 per 1.000 inferenze

Risultato:

- € 0,037 per inferenza
- € 37,00 per 1.000 inferenze

**Confronto sintetico**

Nello Scenario A con API esterna il LCOAI per 1.000 inferenze è pari a € 17,50.

Nello Scenario B con modello self-hosted su infrastruttura dedicata il LCOAI per 1.000 inferenze è pari a € 37,00.

| **Scenario** | **LCOAI €/1.000 inferenze** |
|---|---|
| **API esterna** | € 17,50 |
| **Self-hosted** | € 37,00 |

## 4. Interpretazione per il procurement pubblico

L'esempio mostra che:

Nel breve periodo e con volumi moderati, modelli API risultano economicamente più efficienti.

Il self-hosting diventa competitivo solo con:

- volumi molto elevati,
- riduzione del CAPEX,
- forte esigenza di controllo o sovranità del dato.

LCOAI consente quindi di:

- superare il confronto basato solo sul "prezzo per token";

- integrare CAPEX e OPEX in un'unica metrica comparabile;
- costruire basi d'asta più aderenti alla realtà operativa;
- simulare scenari di crescita dei volumi;
- individuare soglie di convenienza economica.

## 5. Utilizzo nella determinazione della base d'asta

La stazione appaltante può:

- Stimare il volume annuo di inferenze attese.
- Stimare CAPEX e OPEX realistici.
- Calcolare un LCOAI di riferimento.

Utilizzarlo per:

- verificare la coerenza delle offerte,
- evitare sottostime strategiche,
- individuare squilibri tra investimento iniziale e costi ricorrenti.

***N.B.: Le stime riportate sono riferite al contesto e ai prezzi vigenti alla data di pubblicazione del presente allegato e hanno finalità meramente esemplificativa.***