Documentazione LaTeX: Formula Globale per il Calcolo del Costo di un'Applicazione AI

Di seguito viene presentata una **documentazione in LaTeX** che descrive, *nel dettaglio*, la *formula complessiva* per il calcolo del *costo* di un'Applicazione AI basata su:

- **MongoDB Atlas** (database e vector store),
- **AWS Cognito** (autenticazione utenti),
- **OpenAI GPT-4o *vs.* GPT-4o Mini** (LLM per inferenza e embedding),
- **Unstructured.io** (per processare documenti).

La trattazione integra tutte le informazioni delle ricerche precedenti, in particolare sull'uso e i *costi unitari* dei provider, e su come *comporre* il costo totale mensile per utente.

1. Costo Mensile: Parametri, Struttura e Formula Finale

Vogliamo calcolare il **costo mensile** generato da un *singolo utente medio*, sommando diverse componenti:

$$C_{\text{utente,mese}} \;=\; C_{\text{chat}}(\dots) \;+\; C_{\text{ingestion}}(\dots) \;+\; C_{\text{storage}}(\dots)$$

dove:

- C_{chat} rappresenta i costi dovuti all'uso in *chat* (LLM usage),
- $C_{\text{ingestion}}$ i costi di caricamento/indicizzazione di file (documenti, immagini, video),
- C_{storage} l'onere mensile per lo *storage* su *MongoDB Atlas*.

In generale, avremo:

$$\boxed{C_{\text{utente,mese}} \;=\; \underbrace{N_{\text{msg}} \cdot C_{\text{msg}}}_{\text{Chat LLM usage}} \;+\; \underbrace{\sum C_{\text{upload}}}_{\text{Ingestion di documenti/media}} \;+\; \underbrace{S \times c_{\text{GB/mese}}}_{\text{Storage su DB}}}$$

Nel seguito, *dissezioniamo* le singole componenti e i **parametri** che le definiscono.

1.1 Costo Chat: C_{chat}

Consideriamo un utente che *mensilmente* invia N_{msg} messaggi in chat (es. 80). Ciascun *messaggio* (turno utente+risposta AI) ha un costo:

$$C_{\text{msg}} \;=\; C_{\text{LLM}} + C_{\text{retrieval}} + C_{\text{store}}$$

dove:

- $C_{\text{LLM}} = \dfrac{T_{\text{in}}}{1000}\, p_{\text{in}} + \dfrac{T_{\text{out}}}{1000}\, p_{\text{out}}$,
- $C_{\text{retrieval}}$ copre embedding query *(100–200 token?)* e query vector su Atlas ($\sim 10^{-5}$\$)), C_{store} copre scrittura *2 doc* (messaggio+risposta) su DB ($\sim 10^{-6}$–10^{-5}\$)).

Token in Input (T_{in})

$$T_{\text{in}} \;=\; (\#\text{ coppie history} \times 2 \times \bar{t}_{\text{hist}}) \;+\; \bar{t}_{\text{user}} \;+\; (N_{\text{kbox}} \times R_{\text{per\_kbox}} \times \bar{t}_{\text{chunk}})$$

Così $\bar{t}_{\text{hist}}$ = token per messaggio di *storia*, $\bar{t}_{\text{user}}$ = token *input utente* nel messaggio corrente, e $\bar{t}_{\text{chunk}}$ = token medi di un chunk di retrieval. $N_{\text{kbox}}, R_{\text{per\_kbox}}$ = # KBox e chunk per KBox.

Token in Output (T_{out})

$$T_{\text{out}} = \bar{t}_{\text{out}}$$

(es. 200–300 token medi in risposta LLM).

Costi LLM Input/Output Dipendono dal modello scelto:

- GPT-4o (*full*): $p_{\text{in}} = 0.005\$/1k,\ p_{\text{out}} = 0.015\$/1k$,
- GPT-4o *Mini*: $p_{\text{in}} = 0.00015\$/1k,\ p_{\text{out}} = 0.00060\$/1k$.

Retrieval e Store (fissi).

$$C_{\text{retrieval}} \approx 10^{-5}\$, \quad C_{\text{store}} \approx 10^{-5}\$.$$

Il *totale* $\approx (T_{\text{in}} p_{\text{in}} + T_{\text{out}} p_{\text{out}})/1000 + 2 \times 10^{-5}\$$.

1.2 Costo Ingestion: $C_{\text{ingestion}}$

Questo *include* i costi di **caricamento** di *documenti*, *immagini*, *video*. In generale:

$$C_{\text{ingest,mese}} = \sum_{\text{contenuti}} \Big[C_{\text{processing}} + C_{\text{embedding}} + C_{\text{DB}} \Big].$$

Documenti (PDF)

- *Unstructured* =¿ 0.01\$/pagina (Hi-Res) o 0.001\$/pagina (Fast).
- *Chunking* =¿ chunk $\approx T_{\text{doc}}/\bar{T}_{\text{chunk}}$.
- *Embedding doc* =¿ $\frac{\bar{T}_{\text{chunk}}}{1000} \times c_{\text{embed}}$ per chunk.
- *Scrittura DB* =¿ $7.5 \times 10^{-6}\$ per chunk$.

Immagini

- Possibile GPT-4o (o Mini) per *caption* =¿ (token img input + token out).
- Embedding breve descrizione.

Totale $\sim$ (0.002\$GPT-4o or 0.0001\$Mini).

Video

- Estrazione frame =¿ caption LLM + embedding (unificato).
- Se 2 min, sampling 10s =¿ $\sim$ 12 frame =¿ cost LLM $\sim 12 \times cost\_frame$.

1.3 Costo Storage Mensile: C_{storage}

Lo *storage* su **MongoDB Atlas** costa $\sim 0.25\$/GB/mese$. Se l'utente conserva S GB di contenuti (doc, media, embedding):

$$C_{\text{storage,mese}} = S \times c_{\text{GB}}.$$

1.4 Formula Finale

$$\boxed{C_{\text{utente,mese}} = \underbrace{N_{\text{msg}} \cdot C_{\text{msg}}}_{\text{Chat usage}} + \underbrace{\sum_{\text{contenuti caricati}} C_{\text{upload}}}_{\text{Ingestion}} + \underbrace{S \cdot c_{\text{GB/mese}}}_{\text{Storage}} \cdot}$$

2. Dati di Costo Unitari (Provider)

Riportiamo i *costi unitari* principali, come dalle specifiche:

- **MongoDB Atlas**:
  - Lettura RPU: $\sim \$0,10$ / 1M,
  - Scrittura WPU: $\sim \$1,25$ / 1M,
  - Storage: $\sim \$0,25$ / GB-mese.
- **AWS Cognito** (MAU based):
  - 0–50k MAU: gratis,
  - 50k–100k: $\sim \$0,0055$/utente,
  - 100k–1M: $\sim \$0,0046$/utente, ecc.
- **OpenAI GPT-4o**:
  - input: \$0,005/1k,
  - output: \$0,015/1k
- **OpenAI GPT-4o Mini**:
  - input: \$0,00015/1k,
  - output: \$0,00060/1k
- **OpenAI Embedding** (text-embedding-3-small):
  - \$0,00002/1k
- **Unstructured.io**:
  - pipeline Fast: \$0,001/pagina,
  - pipeline Hi-Res: \$0,01/pagina.

3. Parametri e Significato

Elenchiamo **tutti i parametri** (con eventuali valori tipici):

$N_{\mathbf{msg}}$ numero di messaggi di chat dell'utente nel mese (es. 80).

$C_{\mathbf{msg}}$ costo di un singolo messaggio (vedi formula LLM+retr+store).

D, I, V numero documenti, immagini, video caricati al mese.

$c_{\mathbf{page}}$ costo di processare 1 pagina (0,001\$ *Fast* o 0,01\$ *Hi-Res*).

$c_{\mathbf{GB}}$ costo di 1 GB di storage al mese (0,25\$).

S dimensione (GB) dei dati totali dell'utente su Atlas.

$\bar{T}_{\mathbf{in}}$, $\bar{T}_{\mathbf{out}}$ token input e output medi, se usiamo stima fissa (p.es. 1500 in, 300 out).

$\bar{T}_{\mathbf{hist}}$, $\bar{T}_{\mathbf{user}}$ token storia + user.

$p_{\mathbf{in}}, p_{\mathbf{out}}$ costi LLM input/output, *GPT-4o vs GPT-4o Mini.*

α micro-costo retrieval+store (es. 2×10^{-5}).

4. Conclusione e Esempio

Esempio rapido:

- $N_{\text{msg}} = 80$, $C_{\text{msg}} = 0,0025\$$ (mix GPT-4o & Mini),
- $\sum C_{\text{upload}} \approx 0,0004\$$ (poche decine di doc e img),
- $S = 0,1$ GB, $c_{\text{GB/mese}} = 0,25$.

$$C_{\text{utente,mese}} = 80 \times 0,0025 + 0,0004 + 0,1 \times 0,25 = 0,20 + 0,0004 + 0,025 = 0,2254 \approx 0,23\$.$$

Quindi $\sim 0,23\$$ al mese per utente *medio* in questo scenario.

Sostenibilità: Come mostrato, *GPT-4o Mini* riduce drammaticamente i costi di inferenza, e lo *storage* su *Atlas* resta ragionevole. Anche con massima attività (1GB e molte chat) siamo entro 1–2$/utente/mese.