



# Code7Crusaders

Software Development Team

Analisi dei Modelli di linguaggio

## **Membri del Team:**

Enrico Cotti Cottini, Gabriele Di Pietro, Tommaso Diviesti  
Francesco Lapenna, Matthew Pan, Eddy Pinarello, Filippo Rizzolo

**Data:** 17 febbraio 2025

# Indice

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Obiettivo</b>                                 | <b>3</b>  |
| <b>2</b> | <b>Caratteristiche dei Modelli</b>               | <b>3</b>  |
| 2.1      | BigScience Workshop (Huggingface) . . . . .      | 3         |
| 2.1.1    | Prestazioni . . . . .                            | 4         |
| 2.2      | Considerazioni sull'hardware . . . . .           | 4         |
| 2.2.1    | Tabella dei Modelli . . . . .                    | 4         |
| 2.3      | OpenAI . . . . .                                 | 5         |
| 2.3.1    | Prestazioni . . . . .                            | 5         |
| <b>3</b> | <b>Costi e Limiti</b>                            | <b>6</b>  |
| 3.1      | Huggingface . . . . .                            | 6         |
| 3.1.1    | Costi . . . . .                                  | 6         |
| 3.2      | OpenAI . . . . .                                 | 7         |
| 3.2.1    | Costi . . . . .                                  | 7         |
| <b>4</b> | <b>Utilizzo tramite LangChain</b>                | <b>9</b>  |
| 4.1      | BigScience Workshop (Huggingface) . . . . .      | 9         |
| 4.2      | OpenAI . . . . .                                 | 9         |
| <b>5</b> | <b>Testing con OpenAI GPT-4omini e LangChain</b> | <b>10</b> |
| 5.1      | Setup del Test . . . . .                         | 10        |
| 5.2      | Esecuzione del Test . . . . .                    | 10        |
| 5.3      | Conclusioni del Test . . . . .                   | 10        |
| <b>6</b> | <b>Conclusioni</b>                               | <b>11</b> |

## Elenco delle figure

|   |   |    |
|---|---|----|
| 1 | Benchmark delle prestazioni di varie versioni di BLOOM . . . . .                          | 4  |
| 2 | Benchmark delle prestazioni del modello gpt4o confrontato con altri modelli . . . . .     | 5  |
| 3 | Benchmark delle prestazioni del modello gpt4omini confrontato con altri modelli . . . . . | 6  |
| 4 | Costi degli endpoint API di Huggingface per modelli di grandi dimensioni . . . . .        | 7  |
| 5 | Costi delle API di OpenAI per il modello gpt4o . . . . .                                  | 8  |
| 6 | Costi delle API di OpenAI per il modello gpt4omini . . . . .                              | 9  |
| 7 | Scenario riassuntivo dei costi delle varie query su LangSmith . . . . .                   | 10 |

# 1 Obiettivo

Questo documento si pone l'obiettivo di confrontare i modelli di Huggingface, in particolare quelli sviluppati nell'ambito del BigScience Workshop (come BLOOM e le sue varianti), e i modelli di OpenAI (ad esempio GPT-4o e GPT-4omini). Entrambi i tipi di modelli possono essere integrati tramite l'interfaccia LangChain per lo sviluppo di applicazioni avanzate basate su modelli linguistici. Il confronto considera caratteristiche, vantaggi, svantaggi, costi e altri aspetti tecnici rilevanti.

## 2 Caratteristiche dei Modelli

### 2.1 BigScience Workshop (Huggingface)

I modelli di BigScience sono il risultato di un'iniziativa collaborativa open source che mira a democratizzare l'accesso alle tecnologie avanzate di NLP<sup>G</sup>. Tra i modelli più noti figurano BLOOM e le sue versioni ottimizzate come BLOOMz.

#### Caratteristiche principali

- **Open Source:** Il codice sorgente è completamente accessibile, permettendo agli sviluppatori di personalizzare e adattare i modelli alle proprie esigenze.
- **Supporto Multilingue:** I modelli sono progettati per funzionare su una vasta gamma di lingue, incluse molte lingue meno comuni.
- **Dimensioni Variabili:** Sono disponibili modelli di diverse dimensioni, da versioni leggere (adatte a risorse hardware limitate) a modelli complessi che richiedono infrastrutture avanzate.
- **Hosting su Huggingface:** I modelli possono essere utilizzati tramite la piattaforma Huggingface, sia attraverso endpoint API che tramite infrastrutture cloud personalizzate.

#### Vantaggi

- **Accessibilità:** Non ci sono vincoli di licenza proprietaria, il che garantisce un utilizzo flessibile.
- **Trasparenza:** Maggiore chiarezza sui dati di addestramento e sull'architettura del modello.
- **Personalizzazione:** Possibilità di ottimizzare il modello per specifici casi d'uso.

#### Svantaggi

- **Prestazioni Inferiori:** Risultati meno ottimali rispetto ai modelli proprietari in applicazioni altamente specifiche.
- **Requisiti Hardware Elevati:** I modelli più grandi richiedono notevoli risorse computazionali per l'addestramento e l'inferenza.
- **Costo per Grandi Modelli:** L'utilizzo tramite endpoint API è limitato a modelli con dimensioni inferiori a 10GB; per modelli più grandi è necessaria una licenza a pagamento con costi variabili.

### 2.1.1 Prestazioni

Benchmark<sup>G</sup> e valutazione dei modelli di BigScience Workshop dimostrano prestazioni notevoli.



Figure 4: Zero-shot multilingual task generalization with English prompts. BLOOM models have 176 billion parameters. Scores are the language average for each task. Appendix §B breaks down performance by language.

Figura 1: Benchmark delle prestazioni di varie versioni di BLOOM

## 2.2 Considerazioni sull'hardware

Nel caso si volessere utilizzare modelli di grandi dimensioni, come BLOOM, è necessario considerare l'hardware richiesto per l'inferenza. Questi modelli richiedono risorse significative, come GPU ad alte prestazioni e memoria RAM dedicata, per garantire prestazioni ottimali. Nel nostro caso utilizziamo API endpoint che ci permettono di fare affidamento a risorse cloud per l'inferenza, riducendo la necessità di hardware locale, nel caso dei modelli di Huggingface però, gli API endpoint sono gratuiti solo fino a modelli inferiori ai 10GB, per modelli più grandi è necessario sottoscrivere un piano a pagamento.

### 2.2.1 Tabella dei Modelli

Di seguito è riportata una tabella che elenca i modelli di Huggingface e OpenAI con le loro caratteristiche principali:

| Modello     | Parametri       | Dimensione (GB)   | Hardware Necessario                        |
|-------------|-----------------|-------------------|--|
| BLOOM-176B  | 176 miliardi    | 350 GB (FP32)     | 8x NVIDIA A100 80GB o equivalenti          |
| BLOOM-7.1B  | 7,1 miliardi    | 14 GB (FP32)      | 1x NVIDIA A100 40GB o equivalenti          |
| BLOOM-3B    | 3 miliardi      | 6 GB (FP32)       | 1x NVIDIA RTX 3090 (24GB) o equivalenti    |
| BLOOM-1.1B  | 1,1 miliardi    | 2,2 GB (FP32)     | 1x NVIDIA RTX 2080 Ti (11GB) o equivalenti |
| BLOOM-560M  | 560 milioni     | 1,1 GB (FP32)     | 1x NVIDIA GTX 1660 (6GB) o equivalenti     |
| BLOOM-350M  | 350 milioni     | 0,7 GB (FP32)     | 1x GPU integrata con almeno 4GB di memoria |
| BLOOMZ-176B | 176 miliardi    | 350 GB (FP32)     | 8x NVIDIA A100 80GB o equivalenti          |
| GPT-4o      | 70-100 miliardi | 140-200 GB (FP32) | 4-8x NVIDIA A100 80GB o equivalenti        |
| GPT-4omini  | 3-10 miliardi   | 6-20 GB (FP32)    | 1x NVIDIA RTX 3090 (24GB) o equivalenti    |

Tabella 1: Caratteristiche dei modelli di Huggingface e OpenAI

Bisogna anche considerare il fatto che la dimensione dei modelli non è l'unico fattore determinante per le risorse necessarie, ma anche gli eventuali modelli di embedding e i dataset utilizzati influiscono. È quindi necessario prevedere un margine maggiore per l'hardware utilizzato. I modelli piccoli come BLOOM-3B o inferiori sono poco efficaci.

## 2.3 OpenAI

I modelli di OpenAI includono varianti all'avanguardia come GPT-4 e GPT-4omini, sviluppate per fornire elevate prestazioni in una vasta gamma di applicazioni NLP<sup>G</sup>.

### Caratteristiche principali

- **Proprietari:** I modelli sono accessibili esclusivamente tramite API commerciali.
- **Prestazioni Elevate:** OpenAI è leader nei benchmark NLP<sup>G</sup>, garantendo risultati ottimali per applicazioni generative e analitiche.
- **Ottimizzazione per Prodotti Commerciali:** I modelli sono ottimizzati per applicazioni pratiche come chatbot, generazione di codice, automazione aziendale e analisi dei dati.

### Vantaggi

- **Inferenza Scalabile:** Le API cloud di OpenAI permettono una scalabilità elevata senza la necessità di gestione locale delle risorse.
- **Supporto Tecnico:** Documentazione esaustiva e supporto continuo per l'integrazione.
- **Aggiornamenti Regolari:** Miglioramenti costanti ai modelli e alle API.

### Svantaggi

- **Costi Basati sui Token:** I costi dipendono dal numero di token elaborati, il che può risultare oneroso in alcuni scenari ad alto volume.
- **Mancanza di Trasparenza:** Non è possibile accedere ai dati di addestramento o all'architettura interna del modello.
- **Dipendenza dall'Infrastruttura Cloud:** Non è possibile eseguire i modelli localmente, il che può rappresentare un problema per progetti con restrizioni di privacy.

### 2.3.1 Prestazioni

Benchmark<sup>G</sup> e valutazione dei modelli di OpenAI: GPT-4o e GPT-4omini, dimostrano prestazioni superiori rispetto ad altri modelli di riferimento.



Figura 2: Benchmark delle prestazioni del modello gpt4o confrontato con altri modelli



Figura 3: Benchmark delle prestazioni del modello gpt4omini confrontato con altri modelli

### 3 Costi e Limiti

Una differenza chiave tra le due piattaforme riguarda i costi e le limitazioni di utilizzo:

#### 3.1 Huggingface

- Gli endpoint API di Huggingface supportano modelli fino a 10GB gratuitamente.
- Per modelli più grandi è necessario sottoscrivere un piano a pagamento, con costi variabili in base alla dimensione e **all'utilizzo orario**.
- L'utilizzo locale richiede risorse hardware significative, con costi indiretti per l'acquisto o il noleggio di infrastrutture adeguate.

##### 3.1.1 Costi

Nel nostro caso, data la necessità di utilizzare modelli di grandi dimensioni come BLOOM (*a causa dell'inefficienza di quelli piccoli*), i costi di Huggingface potrebbero risultare proibitivi, quindi di seguito mostriamo una tabella che elenca i costi di un API endpoint:



Figura 4: Costi degli endpoint API di Huggingface per modelli di grandi dimensioni

## 3.2 OpenAI

- I costi sono calcolati in base al numero di token utilizzati durante l'elaborazione, rendendo il modello economicamente flessibile per casi d'uso specifici.
- Non ci sono costi orari fissi, e l'utilizzo è facilmente scalabile in funzione delle esigenze del progetto.

Questa differenza rende OpenAI una scelta più conveniente per progetti con utilizzo intermittente o moderato, mentre Huggingface è più adatto a sviluppatori con infrastrutture locali preesistenti o budget elevati per l'acquisto di risorse.

### 3.2.1 Costi

Di seguito i costi di OpenAI per l'utilizzo delle API per gpt4o e gpt4omini, che a differenza di Huggingface non richiedono costi fissi orari ma sono basati sul numero di token utilizzati:



| Model                           | Pricing                           | Pricing with Batch API*   |
|---------------------------------|-----------------------------------|---------------------------|
| gpt-4o                          | \$2.50 / 1M input tokens          | \$1.25 / 1M input tokens  |
|                                 | \$1.25 / 1M cached** input tokens |                           |
|                                 | \$10.00 / 1M output tokens        | \$5.00 / 1M output tokens |
| gpt-4o-2024-11-20               | \$2.50 / 1M input tokens          | \$1.25 / 1M input tokens  |
|                                 | \$1.25 / 1M cached** input tokens |                           |
|                                 | \$10.00 / 1M output tokens        | \$5.00 / 1M output tokens |
| gpt-4o-2024-08-06               | \$2.50 / 1M input tokens          | \$1.25 / 1M input tokens  |
|                                 | \$1.25 / 1M cached** input tokens |                           |
|                                 | \$10.00 / 1M output tokens        | \$5.00 / 1M output tokens |
| gpt-4o-audio-preview            | Text                              |                           |
|                                 | \$2.50 / 1M input tokens          |                           |
|                                 | \$10.00 / 1M output tokens        |                           |
|                                 | Audio***                          |                           |
|                                 | \$100.00 / 1M input tokens        |                           |
| gpt-4o-audio-preview-2024-10-01 | \$200.00 / 1M output tokens       |                           |
|                                 | Text                              |                           |
|                                 | \$2.50 / 1M input tokens          |                           |
|                                 | \$10.00 / 1M output tokens        |                           |
|                                 | Audio***                          |                           |
| gpt-4o-2024-05-13               | \$100.00 / 1M input tokens        |                           |
|                                 | \$200.00 / 1M output tokens       |                           |
|                                 |                                   |                           |
| gpt-4o-2024-05-13               | \$5.00 / 1M input tokens          | \$2.50 / 1M input tokens  |
|                                 | \$15.00 / 1M output tokens        | \$7.50 / 1M output tokens |

Figura 5: Costi delle API di OpenAI per il modello gpt4o

| Model                  | Pricing                            | Pricing with Batch API*   |
|------------------------|------------------------------------|---------------------------|
| gpt-4o-mini            | \$0.150 / 1M input tokens          | \$0.075 / 1M input tokens |
|                        | \$0.075 / 1M cached** input tokens |                           |
|                        | \$0.600 / 1M output tokens         |                           |
| gpt-4o-mini-2024-07-18 | \$0.150 / 1M input tokens          | \$0.075 / 1M input tokens |
|                        | \$0.075 / 1M cached** input tokens |                           |
|                        | \$0.600 / 1M output tokens         |                           |

Figura 6: Costi delle API di OpenAI per il modello gpt4omini

## 4 Utilizzo tramite LangChain

LangChain è una libreria versatile che consente di combinare modelli linguistici avanzati con strumenti di ricerca, database e altre tecnologie. Entrambe le piattaforme possono essere integrate facilmente tramite LangChain.

### 4.1 BigScience Workshop (Huggingface)

LangChain<sup>G</sup> supporta l'integrazione con i modelli Huggingface tramite connettori diretti, permettendo un elevato grado di personalizzazione.

#### Punti di forza

- Possibilità di controllare in dettaglio le modalità di esecuzione del modello.
- Personalizzazione avanzata per applicazioni specifiche.

#### Limiti

- Requisiti tecnici e computazionali elevati per ottimizzare le prestazioni.
- Gestione complessa per l'utilizzo di modelli di grandi dimensioni.

### 4.2 OpenAI

LangChain<sup>G</sup> offre un'integrazione nativa con le API di OpenAI, semplificando l'implementazione e l'uso dei modelli.

#### Punti di forza

- Configurazione immediata e intuitiva.
- Alta scalabilità e tempi di risposta rapidi grazie all'infrastruttura cloud.

#### Limiti

- Dipendenza dalle API proprietarie.
- Costi ricorrenti legati al consumo di token.

## 5 Testing con OpenAI GPT-4omini e LangChain

Per valutare l'efficacia dell'integrazione di OpenAI GPT-4omini tramite LangChain, abbiamo condotto una serie di test. Abbiamo utilizzato LangSmith per monitorare le prestazioni e raccogliere dati dettagliati sull'elaborazione.

### 5.1 Setup del Test

Abbiamo configurato un ambiente di test utilizzando LangChain<sup>G</sup> per interfacciarci con le API di OpenAI. Il setup includeva:

- Un'istanza di LangChain<sup>G</sup> configurata per utilizzare GPT-4omini.
- Monitoraggio delle richieste e delle risposte tramite LangSmith.
- Un istanza RAG (Retrieval-Augmented Generation), realizzata tramite un database vettoriale FAISS<sup>G</sup> e come embedding model BERT<sup>G</sup>(di Huggingface) runnato in locale(2GB VRAM circa compresi dati file esterni vettorializzati) che ci ha permesso di testare llm su contesti ricevuti da file esterni.

### 5.2 Esecuzione del Test

Abbiamo eseguito diversi scenari di test per valutare soprattutto il costo in token dell'utilizzo di GPT-4omini:

| Name          | Input                       | Output                  | Error             | Start Time           | Latency | Dataset | Annotation Queue | Tokens | Cost         | First Token (ms) | Tag |
|---------------|-----------------------------|-------------------------|-------------------|----------------------|---------|---------|------------------|--------|--------------|------------------|-----|
| LangGraph     | ai: ### Qualità del Pr...   | ai: ### Qualità de...   |                   | 10/12/2024, 19:49:33 | 3.87s   |         |                  | 882    | \$0.0002673  | N/A              |     |
| LangGraph     | ai: Dalla descrizione f...  | ai: Dalla descritto...  |                   | 10/12/2024, 19:49:13 | 2.44s   |         |                  | 747    | \$0.00077325 | N/A              |     |
| LangGraph     | ai: Un processo di qu...    | ai: Un processo d...    |                   | 10/12/2024, 19:26:25 | 8.54s   |         |                  | 1,131  | \$0.00030465 | N/A              |     |
| LangGraph     | ai: Ciao! Come posso        | ai: Ciao! Come po...    |                   | 9/12/2024, 22:17:41  | 1.33s   |         |                  | 84     | \$0.0000027  | N/A              |     |
| LangGraph     | ai: Sembra che tu abb...    | ai: Sembra che tu...    |                   | 9/12/2024, 22:01:45  | 1.61s   |         |                  | 97     | \$0.00003525 | N/A              |     |
| LangGraph     | human: ao                   |                         | KeyboardInterrupt | 9/12/2024, 21:56:10  | 10.41s  |         |                  | 111    | \$0.00009665 | N/A              |     |
| LangGraph     | ai: Sembra che tu abb...    | ai: Sembra che tu...    |                   | 9/12/2024, 21:55:59  | 2.22s   |         |                  | 142    | \$0.00004335 | N/A              |     |
| LangGraph     | ai: Ciao! Come posso        | ai: Ciao! Come po...    |                   | 9/12/2024, 21:53:46  | 1.33s   |         |                  | 82     | \$0.00000258 | N/A              |     |
| LangGraph     | ai: Le Leggi di Ham...      | ai: Le Leggi di Ha...   |                   | 9/12/2024, 20:38:04  | 4.48s   |         |                  | 684    | \$0.0002376  | N/A              |     |
| LangGraph     | ai: Il Deuteronomio è L...  | ai: Il Deuteronomi...   |                   | 9/12/2024, 20:37:45  | 4.48s   |         |                  | 670    | \$0.0002355  | N/A              |     |
| LangGraph     | ai: Il "Messale Roman...    | ai: Il "Messale Ro...   |                   | 9/12/2024, 19:58:07  | 4.58s   |         |                  | 375    | \$0.00079125 | N/A              |     |
| LangGraph     | ai: Il brano che hai for... | ai: Il brano che ha...  |                   | 9/12/2024, 19:54:58  | 7.40s   |         |                  | 872    | \$0.0003558  | N/A              |     |
| LangGraph     | ai: Il testo che hai fom... | ai: Il testo che ha...  |                   | 9/12/2024, 19:54:42  | 8.78s   |         |                  | 833    | \$0.00034995 | N/A              |     |
| LangGraph     | ai: Il testo che hai fom... | ai: Il testo che ha...  |                   | 9/12/2024, 19:52:39  | 7.23s   |         |                  | 1,449  | \$0.00044235 | N/A              |     |
| LangGraph     | ai: Il brano che hai cit... | ai: Il brano che ha...  |                   | 9/12/2024, 19:52:30  | 6.73s   |         |                  | 874    | \$0.0003561  | N/A              |     |
| LangGraph     | ai: La figura dell'om...    | ai: La figura dell'u... |                   | 9/12/2024, 19:52:17  | 5.93s   |         |                  | 1,184  | \$0.0004026  | N/A              |     |
| LangGraph     | ai: Sembra che tu stia...   | ai: Sembra che tu...    |                   | 9/12/2024, 19:51:02  | 8.49s   |         |                  | 576    | \$0.0003914  | N/A              |     |
| LangGraph     | ai: Sembra che tu stia...   | ai: Sembra che tu...    |                   | 9/12/2024, 19:50:40  | 8.22s   |         |                  | 581    | \$0.0003728  | N/A              |     |
| trim_messages | human: Yes, I'm seen...     | human: Yes, I'm s...    |                   | 9/12/2024, 19:46:32  | 0.42s   |         |                  | 0      | N/A          | N/A              |     |

Figura 7: Scenario riassuntivo dei costi delle varie query su LangSmith

### 5.3 Conclusioni del Test

I test hanno dimostrato che l'integrazione di GPT-4omini tramite LangChain<sup>G</sup> è efficace e offre prestazioni soddisfacenti per le nostre esigenze. Il monitoraggio tramite LangSmith ha fornito dati utili per ottimizzare l'uso del modello e gestire i costi associati all'utilizzo dei token.

In conclusione, l'uso di OpenAI GPT-4omini con LangChain<sup>G</sup> rappresenta una soluzione valida per il nostro progetto, garantendo un equilibrio tra prestazioni elevate e costi gestibili.

## 6 Conclusioni

La scelta tra Huggingface e OpenAI dipende dalle esigenze specifiche del progetto e dalle risorse disponibili.

Huggingface è ideale per sviluppatori che richiedono trasparenza, controllo e la possibilità di eseguire modelli su infrastrutture locali. Tuttavia, i costi associati a modelli di grandi dimensioni e le elevate richieste hardware possono rappresentare uno svantaggio.

OpenAI è la soluzione preferibile per applicazioni commerciali o progetti che richiedono elevate prestazioni con un'infrastruttura cloud scalabile e tempi di implementazione rapidi. I costi basati sui token offrono maggiore flessibilità economica rispetto ai costi fissi di Huggingface.

Grazie a LangChain, entrambe le opzioni possono essere integrate efficacemente, permettendo di sfruttare appieno le potenzialità dei modelli linguistici per applicazioni avanzate.

Tuttavia, per il nostro progetto **Capitolato 7 Ergon**, l'opzione con OpenAI sembra essere la più adatta all'implementazione pratica, considerati i nostri mezzi e possibilità.

Data: \_\_\_\_\_

Firma: \_\_\_\_\_