

M.INC.

Dipartimento di Ricerca e Sviluppo

Rapporto Benchmark AI – Qualità Generazione Testi

Autore: Mattimax

Data: 16 Maggio 2025

1. Introduzione

Il presente rapporto analizza le prestazioni della quarta generazione della nostra linea di modelli AI, **DATA-AI_Chat_4_0.6B_Q8**, sviluppata da M.INC., ponendola a confronto con altri modelli di riferimento nella fascia sotto i 2B di parametri.

Questa nuova versione segna un netto miglioramento rispetto alle precedenti iterazioni, in termini di **espressività, fluidità e attinenza al prompt**, pur mantenendo **dimensioni contenute (600M parametri)** e **alta compatibilità su sistemi CPU-only**.

2. Metodologia

2.1 Modelli Confrontati

I modelli selezionati per il benchmark includono:

- DATA-AI_Chat_4_0.6B_Q8** (M.INC.)
- Lite-Oute-1-300M**
- Deepseek-R1_1.5B**
- Qwen2.5-0.5B**
- Llama-3.2-1B**

Questi modelli sono stati testati su un dataset comune con prompt in italiano, coprendo una varietà di domini (narrativo, tecnico, informativo).

2.2 Parametri di Valutazione

Sono stati valutati due insiemi di metriche:

A. Valutazione linguistica e stilistica

(Scala 1-10):

- Grammatica
- Lessico
- Coerenza
- Stile Narrativo
- Espressività

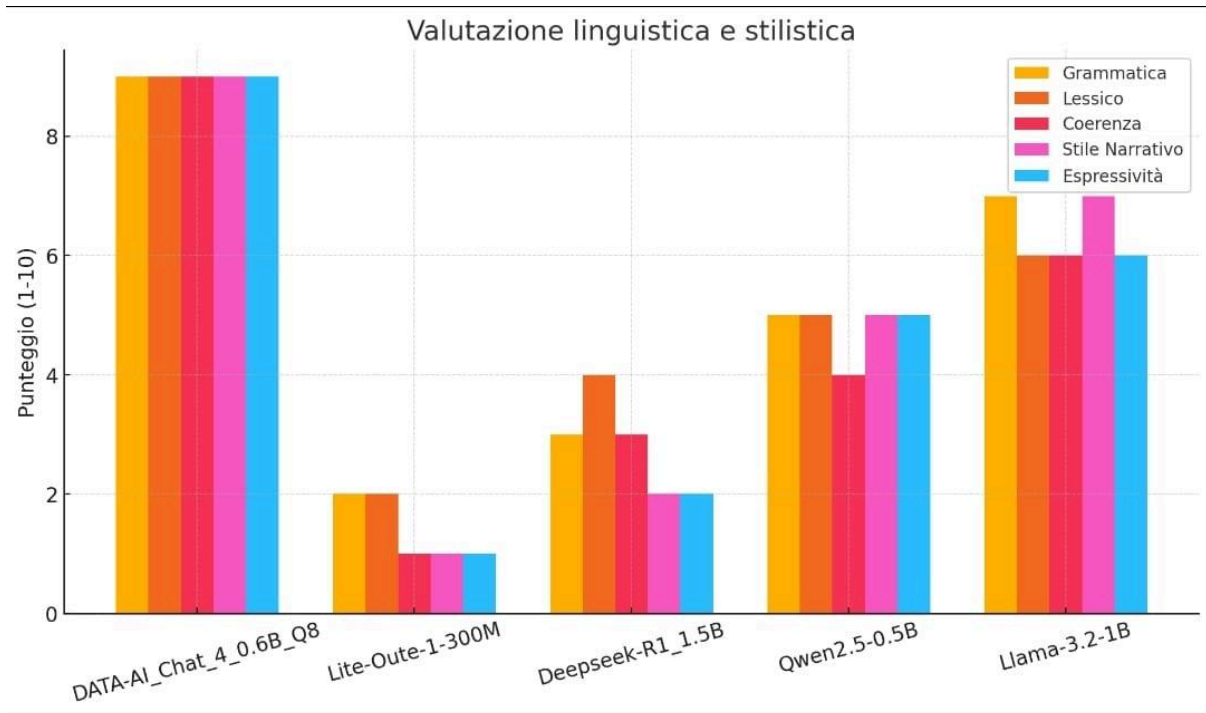
B. Risposta al Prompt e Contenuto

(Scala 1-10):

- Attinenza
 - Originalità
 - Fluidità
 - Emozione
 - Complessità
-

3. Risultati e Analisi

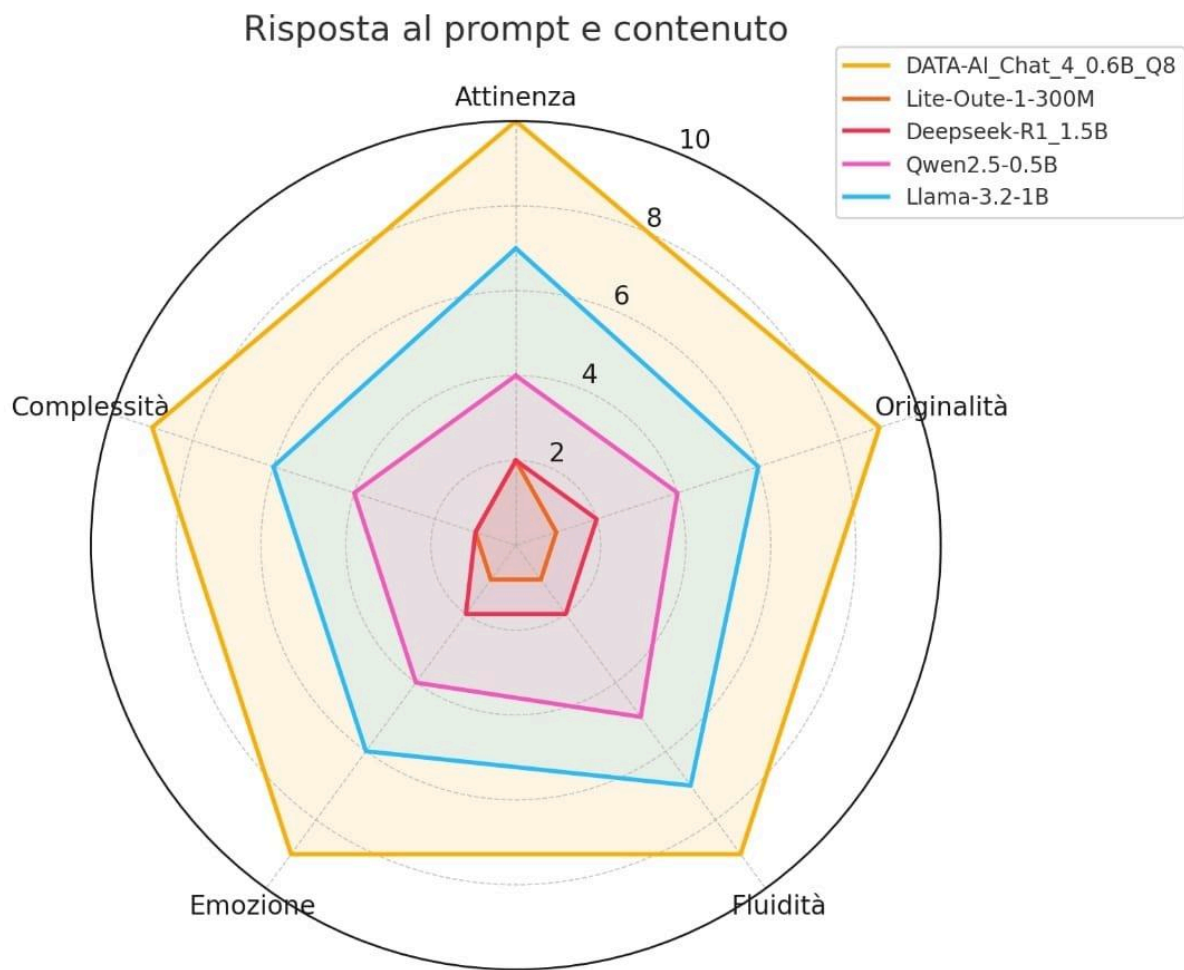
3.1 Grafico a Barre – Linguaggio e Stile



Osservazioni principali:

- **DATA-AI_Chat_4_0.6B_Q8** eccelle in tutte le metriche linguistiche, superando **9/10** in grammatica, lessico, coerenza e stile.
- Gli altri modelli mostrano performance inferiori, specialmente **Lite-Oute-1-300M**, che fatica a superare il 2.
- **Llama-3.2-1B** e **Qwen2.5-0.5B** mostrano prestazioni intermedie ma stabili, con valori tra 6 e 7.5.

3.2 Grafico Radar – Pertinenza al Prompt



Osservazioni principali:

- **DATA-AI_Chat_4_0.6B_Q8** domina in **attinenza, originalità e emozione**, con punteggi prossimi al massimo (9.5-10).
 - Si evidenzia una **notevole capacità narrativa e stilistica**, con risposte che risultano **più profonde e coinvolgenti** rispetto ai modelli anche 2-3 volte più grandi.
 - **Llama-3.2-1B** risulta l'unico modello con una forma radar comparabile, ma leggermente inferiore in emozione e complessità.
 - **Deepseek-R1_1.5B** e **Lite-Oute-1-300M** risultano significativamente meno espressivi.
-

4. Conclusioni e Raccomandazioni

Il modello **DATA-AI_Chat_4_0.6B_Q8** si dimostra **nettamente superiore** alla media dei modelli della stessa fascia e anche di alcune versioni più grandi.

Le sue **risposte sono più naturali, creative e attinenti**, mantenendo una **lingua corretta e un tono narrativo coinvolgente**.

Punti di forza:

- Altissimo punteggio in **attinenza, emozione e originalità**.
- Linguaggio scorrevole, corretto, con **stile narrativo maturo**.
- Ottimizzazione eccellente per **CPU-only e dispositivi embedded**.

Raccomandazioni future:




- Esplorare varianti per compiti specifici (es. coding, logica, riassunti).
- Introdurre modalità istruzione controllata e parametri per tonalità emotiva.
- Ottimizzazione ulteriore del modello per input lunghi.

Il progetto può considerarsi **pronto per una release pubblica matura**, e rappresenta un **punto di riferimento** tra i modelli leggeri in italiano.

5. Informazioni Tecniche

- Architettura di base: **DATA-AI_Chat_4_0.5B-Thinking con ristrutturazione intermodulare**
 - Parametri: **600M**
 - Training: **Dataset italiano potenziato + fine-tuning con LoRA**
 - Formati: FP16, GGUF (Q8), compatibilità con llama.cpp, Ollama, ...
-

6. Collegamenti e Risorse

-  [Modello FP16 su HuggingFace](#)
-  [Modello GGUF Q8 per inferenza offline](#)
-  [Deploy diretto con Ollama](#)

Sviluppato con cura da M.INC. – Mattimax
