Massive Text Processing locale con LLM open su CPU

Relatore: **Nicola Ranaldo**

Linux Day 2025

di cosa parliamo

- trattamento massivo di documenti con AI (riassunti, estratti, analisi, classificazione, traduzione, adattamento stile e pubblico, retrieving, generazione, etc.)
- utilizzo di software Libero e modelli OpenWeights in locale con risorse limitate
- verifica della qualità delle elaborazioni su un dominio specifico: atti amministrativi pubblici
- riflessioni

Motivazioni

- l'AI può supportare efficacemente processi e persone nella vita lavorativa e quotidiana
- gli LLM sono utilissimi nelle organizzazioni che trattano documenti, in particolare **Enti Pubblici**
- passione ispirata dalla fantascienza
- software libero
- privacy/lock-in

Privacy e consapevolezza

- tendenza a "raccontare" informazioni personali sensibili
- i provider utilizzano le nostre chat per addestrare i modelli
- i piani free spesso non permettono opt-out
- le chat temporanee spesso vengono conservate per un periodo di tempo e poi eliminate
- fiducia nei provider
- corsa per la supremazia tecnologica con allentamento di controlli
- regolamentazione in corso d'opera
- utilizzo diffuso e non ufficiale nelle organizzazioni

Modelli Open-Weights

- Modelli **open-weight** = pesi pubblicati per esecuzione autonoma
- possibilità di "fine-tuning"
- **Sfida risorse**: memoria, quantizzazione, orchestrazione per su hardware limitato.
- **Batching** : rinunciare alla chat istantanea multi-turn per gestire code di richieste senza interazione.

Il progetto

- Apprendimento della tecnologia per sviluppare soluzioni e capacità progettuali
- Privacy e compliance normativa come prerequisito
- Test bench continuo per valutare modelli, pipeline, prompt...

~~KISS~~


Terraform


Ansible


Kubernetes


FluxCD

Helm


Knative


PostgreSQL

PostgREST


Redis


MinIO

SOPS


Python


Prefect


Keycloak


OpenLDAP

OIDC Guard


GitHub


GitHub Actions


LangChain


Weaviate

Open WebUI

Apache Tika


Ollama

Speaches

JavaScript


Next.js

Orchestrazione

- sviluppo di codice in python (no automations)
- gestione di elaborazioni, scheduling, concorrenza e code con prefect (kubernetes worker)

Prefect Console

Benchmarks

- generici
- non sempre veritieri
- non adatti a domini specifici

Risultati

- circa 4000 elaborazioni in poche settimane
- utilizzo di 8 modelli OpenWeights
- analisi complete su specifici dataset ed a campione (statisticamente non rilevanti)
- confronto con modelli Open Weights di grosse dimensioni (API)
- confronto con modelli proprietari (OpenAI)

Total Failures

The keyword “CIBI” appears on the line that begins with ``` CIBI: ``` —i.e. the line that follows the signature block and serves as the heading for the CIBI-related items.

I’m sorry, but I’m not sure what you would like me to do with this document. Could you let me know whether you need a summary, a translation into another language, or something else?

Errori di anonimizzazione

La Provincia di Benevento concede alla ditta **PINCO PALLINO**, C.F. **…omissis…** l’uso di un pozzo a ...omissis... per irrigazione, con portata massima 0,59 l/s e volume annuo 3 072 m³. Il canone annuale di 54,95 € deve essere versato entro 31 gennaio; è obbligatorio anche se l’acqua non viene utilizzata, salvo rinuncia. Ogni anno può essere rivalutato in base all’inflazione regionale. Il concessionario deve monitorare portata, livello piezometrico e trasmettere dati annuali; inadempimento di tre annualità di canone porta alla decadenza. La concessione è valida 30 anni e può essere revocata dalle autorità di bacino.

Formato/Lingua errati

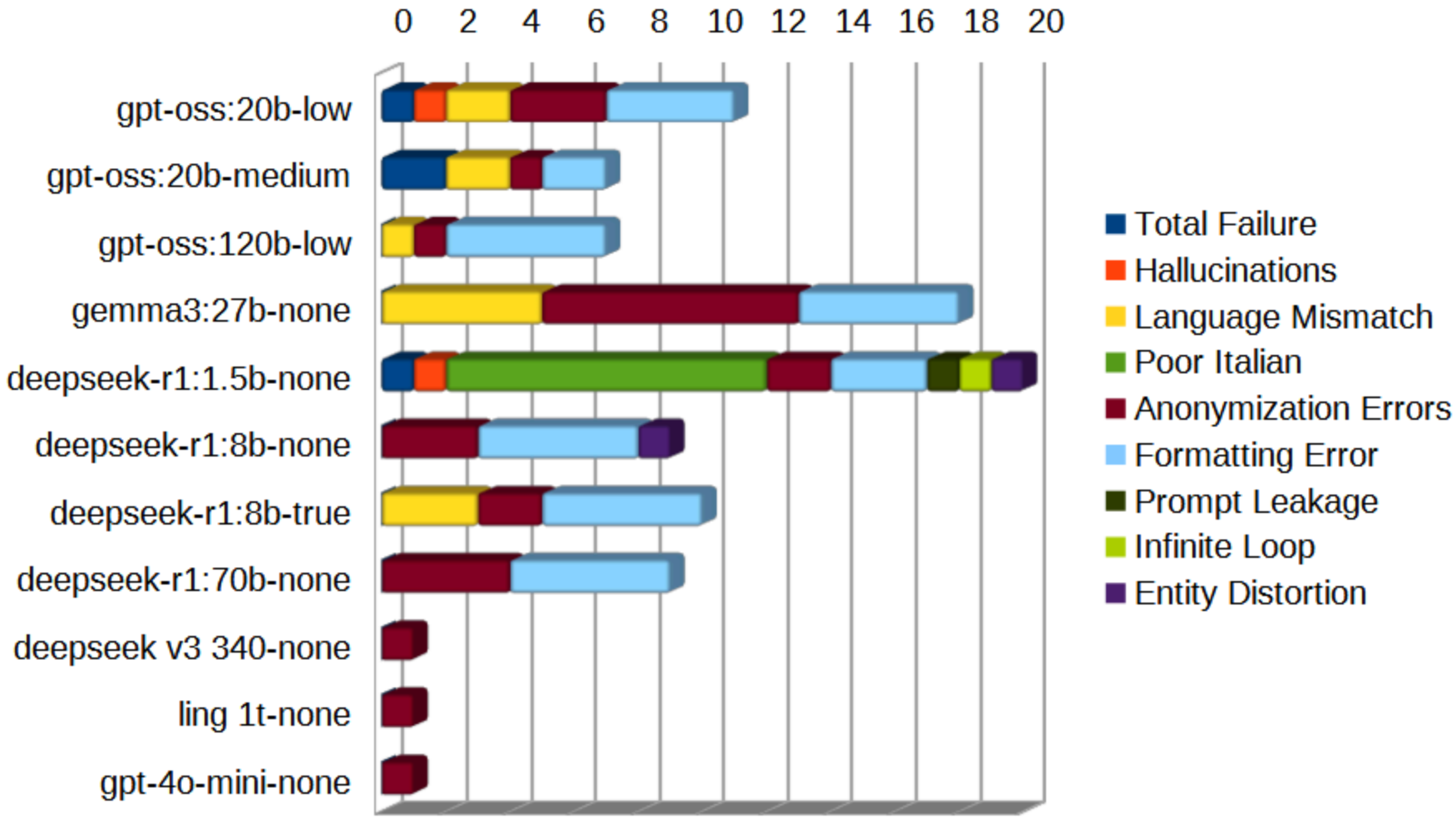
```
- Open a digital tender to rebuild the former school in
...omissis... as a nursery under PNNR Next Generation EU.
- Base contract ...omissis..., managed via Provincia di
Benevento’s e-procurement portal.
- ...omissis...
- Commitments: €250 to ANAC ...omissis...
```

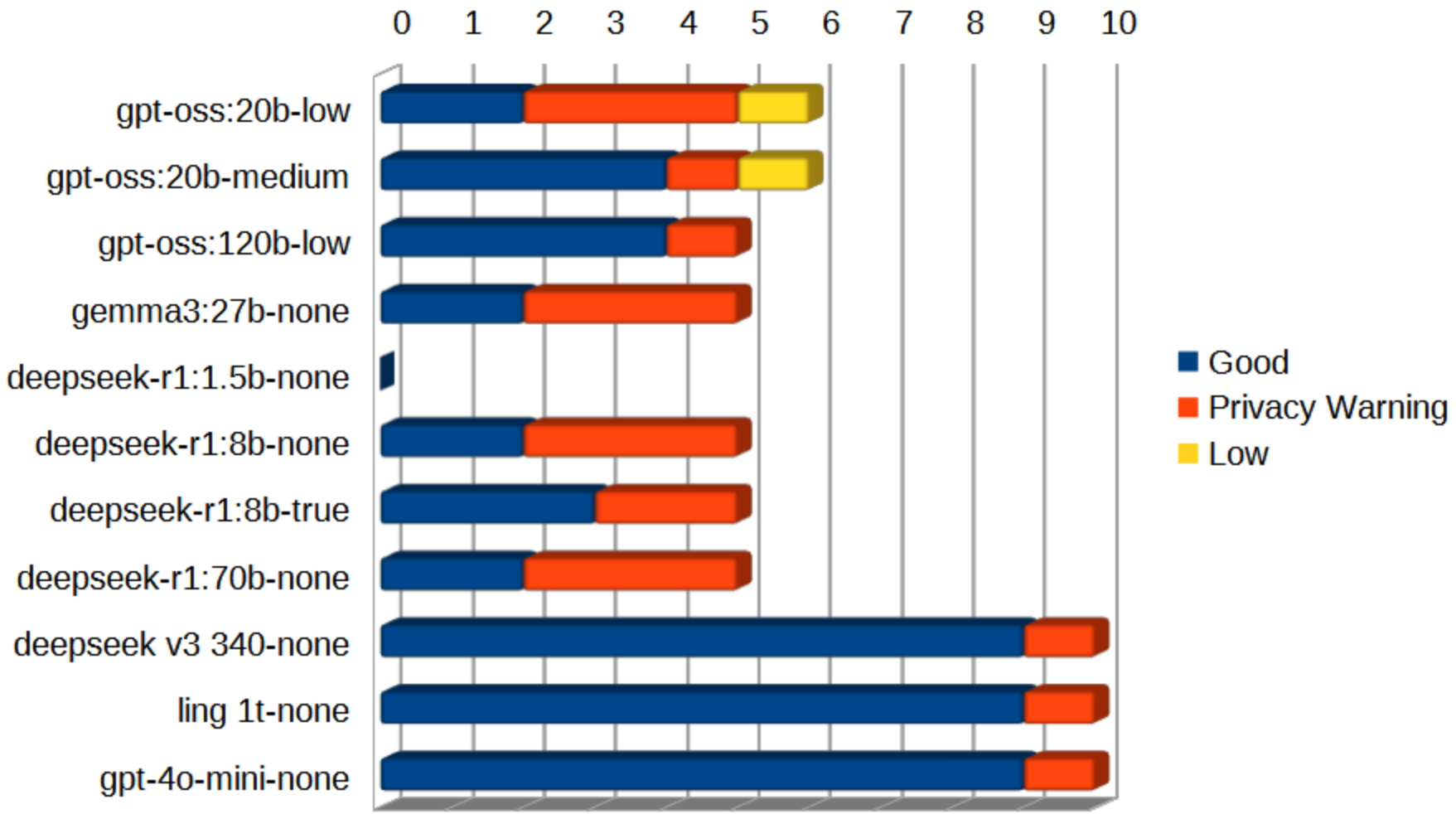
Prompt leaking

```
Leggi il TESTO_ATTO e genera una sintesi in ITALIANO (usando
al massimo 100 parole) che: - sia comprensibile da un ragazzo
di 20 anni - evidenzi effetti gjuridici ed economici, dando
priorità alla parte dispositiva (obblighi, diritti, importi,
termini, condizioni, garanzie, penali, scadenze, percentuali)
- escluda premesse, motivazioni, riferimenti generics a
pubblicazione/trasparenza/correttezza amministrative e
trasmissionta alla ragioneria/settore finanziario per gli
adempimenti consequentiali - anonimizzi, in ottemperanza alla
normativa sulla privacy...
```

infinite loop/poor italian

L'uso di contesti ambientali, come il mancante legame di messaggio tra i si roominess e gli altri ambienti ambientali, contribue a produrre errori di privacy. 5. **Speranza aroundurne in Assicurazioni:** Sospensioni tra si roominess roominess potrebbero essere sospese anche da autori e portarli ad invoglio di soluzioni per risolvere i problemi ambientali. 6. **Suffice di Risorse polizia:** Le condizioni ambientali roominess...


0
2
4
6
8
10
12
14
16
18
20
gpt-oss:20b-low
gpt-oss:20b-medium
gpt-oss:120b-low
gemma3:27b-none
deepseek-r1:1.5b-none
deepseek-r1:8b-none
deepseek-r1:8b-true
deepseek-r1:70b-none
deepseek v3 340-none
ling 1t-none
gpt-4o-mini-none
Total Failure
Hallucinations
Language Mismatch
Poor Italian
Anonymization Errors
Formatting Error
Prompt Leakage
Infinite Loop
Entity Distortion


0
1
2
3
4
5
6
7
8
9
10
gpt-oss:20b-low
gpt-oss:20b-medium
gpt-oss:120b-low
gemma3:27b-none
deepseek-r1:1.5b-none
deepseek-r1:8b-none
deepseek-r1:8b-true
deepseek-r1:70b-none
deepseek v3 340-none
ling 1t-none
gpt-4o-mini-none
Good
Privacy Warning
Low

stima ti tempi e costi

Modello (Reasoning)	Durata 10 atti	Throughput/h	Giorni (24/7)	% tempo a CPU piena (annuo)	Δ costo energia 1× (€/anno)	Δ costo energia ×5 (€/anno)
DeepSeek-R1 70B (none)	4 h 28	2,24	55,8	15,30%	€ 65,33	€ 326,65
Gemma 3 27B (none)	2 h 27	4,08	30,6	8,39%	€ 35,83	€ 179,15
GPT-OSS 120B (low)	2 h 06	4,76	26,3	7,19%	€ 30,71	€ 153,55
DeepSeek-R1 8B (true)	2 h 02	4,92	25,4	6,96%	€ 29,74	€ 148,70
GPT-OSS 20B (medium)	1 h 27	6,90	18,1	4,97%	€ 21,21	€ 106,05
DeepSeek-R1 8B (none)	1 h 25	7,06	17,7	4,85%	€ 20,72	€ 103,60
DeepSeek-R1 1.5B (none)	58 min	10,34	12,1	3,31%	€ 14,14	€ 70,70
GPT-OSS 20B (low)	52 min	11,54	10,8	2,97%	€ 12,68	€ 63,40
GPT-4o mini (none)	46 s	765,96	0,16	—	€ 0,19	€ 0,95

l'IA può commettere errori → la normativa riconosce il rischio e impone controlli e responsabilità.

Legge 23 settembre 2025, n. 132 – "Disposizioni e deleghe al Governo in materia di intelligenza artificiale"

in vigore dal 10 ottobre 2025

L. 132/2025

- Antropocentrico — l’IA al servizio delle persone - tutela dei diritti fondamentali - sviluppo responsabile
- Trasparenza — obbligo di comunicare l’impiego di sistemi automatizzati e garantire tracciabilità
- Sicurezza e gestione del rischio - misure tecnico-organizzative proporzionate
- Supervisione umana e responsabilità - l’IA è prevista come supporto, non come sostituto del giudizio umano
- Aggravanti per reati commessi tramite IA

Grazie!

powered by