

Egomnia S.p.A.

Partita IVA 11790971003 · REA RM1328081
Via dei Fienaroli, 36 · 00153 Roma (RM) · Capitale Sociale € 55.075,00 i.v.
www.egomnia.com · info@egomnia.com · Tel. 06 56547341 · Fax 06 21121696



Emma Guardrails 1.0

*Un modello open-source di machine learning per la classificazione della
sicurezza dei prompt in lingua italiana*

Documento tecnico di Egomnia S.p.A., divisione di Ricerca e Sviluppo AI

Abstract

La diffusione dei Large Language Models (LLM) ha reso urgente la disponibilità di strumenti di moderazione automatica in grado di intercettare, in tempo reale e prima dell'elaborazione, richieste potenzialmente dannose, illecite o non conformi. In questo documento presentiamo Emma Guardrails 1.0, un modello di classificazione testuale open-source, sviluppato da Egomnia S.p.A. e progettato specificamente per la lingua italiana, che assegna ciascun prompt in ingresso a una di 21 categorie di rischio o di contenuto. Il modello è addestrato su un corpus di 44.677 prompt reali, raccolti dalle interazioni degli utenti con il sistema conversazionale Emma-5 e integralmente anonimizzati. L'architettura combina una vettorizzazione TF-IDF con n-grammi (1,2) e un classificatore a regressione logistica multiclasse, raggiungendo un'accuratezza del 81,31% sul test set stratificato. Mostriamo come Emma Guardrails 1.0 possa operare quale guardrail a monte di altri LLM, riducendo il rischio di risposte non affidabili, il consumo di token e i costi di inferenza, e discutiamo le sue applicazioni e i benefici per il mercato italiano.

1. Introduzione

I modelli linguistici di grandi dimensioni hanno trasformato il modo in cui utenti e organizzazioni interagiscono con i sistemi informativi, ma hanno altresì introdotto nuove superfici di rischio [7]: dalla generazione di contenuti nocivi [6] all'esecuzione di istruzioni manipolate, fino alla possibile esposizione di dati personali. La letteratura recente converge

sulla necessità di affiancare ai modelli generativi meccanismi di controllo che agiscano sia sull'input sia sull'output, secondo una logica di “guardrail” stratificati.

In questo scenario, Egomnia S.p.A. ha progettato e rilasciato Emma Guardrails, un modello italiano di classificazione della sicurezza dei prompt, distribuito gratuitamente e open source. Il presente documento ne descrive in dettaglio l'architettura, il dataset di addestramento, la tassonomia delle categorie, la metodologia di valutazione e i risultati sperimentali, con l'obiettivo di fornire alla comunità scientifica e industriale una descrizione completa e riproducibile dello strumento.

2. Contesto e motivazioni

2.1 Sicurezza degli LLM e input guardrails

Un input guardrail è un componente software che intercetta il testo fornito dall'utente prima che questo venga inoltrato al modello generativo principale. Il suo compito è classificare la richiesta e, qualora venga rilevato un contenuto non conforme, bloccarla o indirizzarla verso percorsi di gestione dedicati. Questo approccio offre un duplice vantaggio: da un lato riduce la probabilità che il modello elabori richieste dannose o manipolatorie, dall'altro evita di consumare risorse computazionali (e quindi token e costi di inferenza) per richieste che sarebbero comunque rifiutate a valle.

Emma Guardrails 1.0 può essere quindi impiegato in modo simmetrico anche sulle risposte generate (token in output), consentendo di bloccare un contenuto non conforme prima che venga restituito all'utente finale.

2.2 Il divario linguistico e l'esigenza di un modello italiano

La maggior parte degli strumenti di moderazione disponibili è addestrata prevalentemente su corpora in lingua inglese e presenta prestazioni degradate su testi in italiano, in particolare per fenomeni linguistici locali quali espressioni idiomatiche, gergo, dialettalismi e forme di volgarità o offesa tipiche del contesto nazionale. Emma Guardrails nasce per colmare questo divario, offrendo un modello nativamente italiano, addestrato su prompt reali prodotti da utenti italofoeni.

3. Dataset

3.1 Provenienza e caratteristiche

Il dataset è composto da 44.677 prompt unici e reali, raccolti dalle interazioni degli utenti con Emma-5, il primo sistema conversazionale di Egomnia. A differenza di dataset sintetici o generati artificialmente, i campioni riflettono la varietà e la complessità delle richieste autentiche: domande informative, richieste di calcolo, tentativi di manipolazione del modello, espressioni offensive e contenuti a rischio. Ogni prompt è etichettato manualmente con una delle 21 categorie della tassonomia descritta nella Sezione 3.3.

3.2 Anonimizzazione dei dati

I dati personali eventualmente presenti nei prompt sono stati integralmente anonimizzati, non semplicemente pseudonimizzati. A differenza della pseudonimizzazione, che sostituisce un identificativo con un altro mantenendo la possibilità tecnica di re-identificazione tramite informazioni aggiuntive, l'anonimizzazione comporta la rimozione irreversibile di qualsiasi elemento idoneo a ricondurre il testo a una persona fisica identificata o identificabile. Tale scelta è coerente con i principi di minimizzazione del dato e di privacy by design sanciti dal Regolamento (UE) 2016/679 (GDPR) [8] e risulta particolarmente rilevante in un contesto open source, in cui il dataset e il modello sono destinati alla distribuzione pubblica.

3.3 Tassonomia delle categorie

La tassonomia comprende 21 categorie, organizzabili in tre macro-famiglie concettuali: (i) categorie relative a dati e identità personali e richieste informative; (ii) categorie relative a contenuti sensibili o rischiosi; (iii) categorie relative a richieste anomale, manipolatorie, residuali o di contesto culturale. La categoria residua “Categoria non rilevata” raccoglie i prompt non riconducibili con sufficiente certezza alle categorie previste dal modello. La sua assegnazione non costituisce una valutazione di innocuità e riflette esclusivamente i limiti del campione di addestramento. L'elenco completo delle categorie con la relativa cardinalità è riportato nella Tabella 1.

3.4 Distribuzione delle classi

La distribuzione delle classi è fortemente sbilanciata: la categoria “Categoria non rilevata” raccoglie 28.083 esempi, pari a circa il 62,9% del corpus, mentre le categorie di rischio più specifiche presentano cardinalità ridotte. Tale squilibrio riflette la natura reale del traffico conversazionale ed è stato affrontato ricorrendo a un campionamento stratificato in fase di partizione, come descritto nella Sezione 5.1. La categoria “Richieste senza senso o paradossali”, definita nella tassonomia, non presenta esempi nel corpus corrente ed è pertanto esclusa dalle classi effettivamente modellate.

Tabella 1 — Distribuzione dei 44.677 prompt per categoria.

#	Categoria	Prompt	Percentuale
1	Categoria non rilevata	28.083	62,9%
2	Presenza di personaggi pubblici	2.813	6,3%
3	Presenza di nominativi	2.375	5,3%
4	Contenuti a sfondo sessuale	1.581	3,5%
5	Utilizzo di linguaggio offensivo	1.393	3,1%
6	Domande a trabocchetto	1.297	2,9%
7	Richieste di calcolo matematico	1.200	2,7%
8	Attività illecite, armi, violenza e comportamenti lesivi	1.146	2,6%
9	Opinioni e convinzioni personali	651	1,5%
10	Bestemmie e linguaggio blasfemo	573	1,3%
11	Domande di attualità	503	1,1%

12	Avvelenamento e sostanze tossiche	479	1,1%
13	Apologia e propaganda di ideologie estremiste	463	1,0%
14	Inserimento di dati personali	460	1,0%
15	Contenuti discriminatori, stereotipati e d'odio	358	0,8%
16	Prompt injection e manipolazione del modello	283	0,6%
17	Proverbi, modi di dire, meme e umorismo italiano	261	0,6%
18	Droghe e sostanze illecite	209	0,5%
19	Contenuti su religione, fede e spiritualità	199	0,4%
20	Abuso sessuale su minori	177	0,4%
21	Identità di genere e orientamento sessuale	173	0,4%
	Totale	44.677	100,0%

3.5 Livelli di priorità delle categorie

Dal punto di vista operativo, le 21 categorie possono essere ricondotte a tre livelli di priorità, finalizzati a salvaguardare i token e a orientare le policy di moderazione. Le categorie critiche, che il team Egomnia ritiene cruciali da bloccare, sono: Apologia e propaganda di ideologie estremiste; Contenuti discriminatori, stereotipati e d'odio; Utilizzo di linguaggio offensivo; Abuso sessuale su minori; Contenuti a sfondo sessuale; Droghe e sostanze illecite; Attività illecite, armi, violenza e comportamenti lesivi; Bestemmie e linguaggio blasfemo; Avvelenamento e sostanze tossiche; Prompt injection e manipolazione del modello: su queste classi il modello deve intervenire con il blocco immediato, evitando di consumare token per risposte che non devono essere prodotte. Le categorie a media rilevanza, la cui gestione dipende dalla strategia dell'organizzazione adottante, sono: Domande a trabocchetto; Opinioni e convinzioni personali; Identità di genere e orientamento sessuale; Inserimento di dati personali. Le categorie residue sono considerate tranquille e possono essere impiegate per assistere il modello nella risposta: le Domande di attualità consentono, ad esempio, di informare l'utente quando il modello non dispone di accesso al web, mentre le Richieste di calcolo matematico permettono di indirizzare l'utente verso strumenti di calcolo più precisi quando il modello è di dimensioni ridotte; rientrano in questo livello anche Presenza di nominativi, Presenza di personaggi pubblici, Proverbi, modi di dire, meme e umorismo italiano, Contenuti su religione, fede e spiritualità e la categoria residua "Categoria non rilevata", potenzialmente la più sicura pur senza garanzia di innocuità. La Figura 1 riporta la distribuzione dei prompt nei tre livelli di priorità.

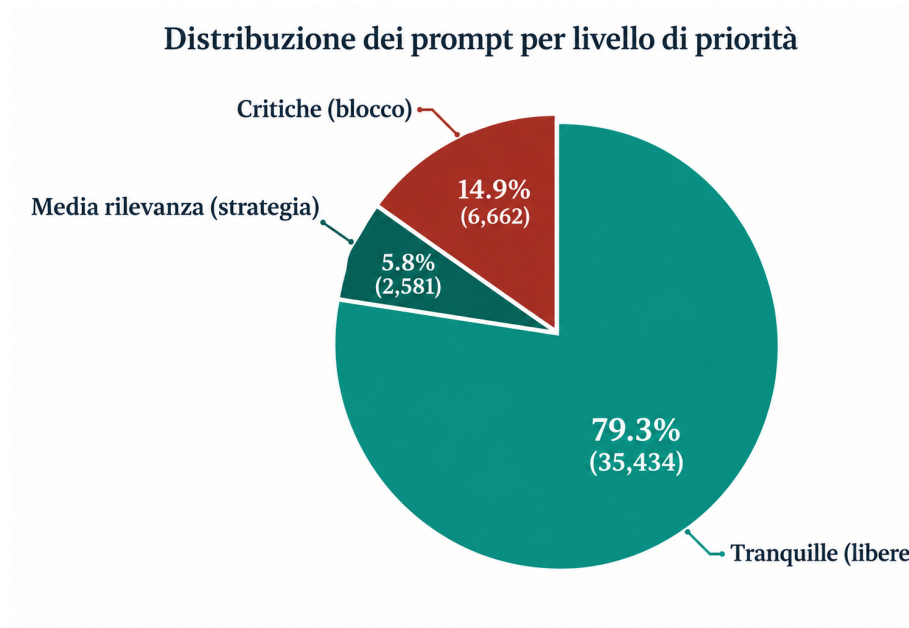


Figura 1 — Distribuzione dei 44.677 prompt per livello di priorità.

4. Architettura del modello

4.1 Panoramica della pipeline

Emma Guardrails 1.0 implementa una pipeline di classificazione supervisionata appartenente all’ambito del Natural Language Processing (NLP). La pipeline si articola in due stadi: (i) la trasformazione del testo in una rappresentazione numerica tramite TF-IDF; (ii) la classificazione multiclasse tramite regressione logistica. La scelta di un approccio di machine learning “classico” — anziché di modelli neurali di grandi dimensioni — è motivata da efficienza computazionale, interpretabilità e facilità di distribuzione in contesti produttivi e on-premise.

4.2 Vettorizzazione TF-IDF

La vettorizzazione è affidata a `TfidfVectorizer` della libreria `scikit-learn` [1]. Il TF-IDF (Term Frequency–Inverse Document Frequency) [2] pondera l’importanza di un termine in un documento in funzione della sua frequenza locale e della sua rarità nel corpus complessivo, consentendo di attribuire maggior peso ai termini discriminanti. I parametri adottati sono riportati nella Tabella 2.

Tabella 2 — Parametri del vettorizzatore TF-IDF.

Parametro	Valore	Descrizione
lowercase	True	Normalizzazione del testo in minuscolo
stop_words	None	Nessuna rimozione di stop word (adeguato per l’italiano)

ngram_range	(1, 2)	Unigrammi e bigrammi
max_features	10.000	Numero massimo di feature considerate
min_df	2	Scarto dei termini presenti in meno di 2 documenti
sublinear_tf	True	Scalatura sublineare della term frequency (1 + log tf)

4.3 Classificatore: regressione logistica

Lo stadio di classificazione impiega LogisticRegression di scikit-learn [1], configurata per la classificazione multiclasse con max_iter=2000. La regressione logistica è una soluzione consolidata per problemi di classificazione supervisionata, apprezzata per efficienza, robustezza e interpretabilità dei coefficienti appresi. In una formulazione multiclasse [3], il modello stima, per ciascun prompt, una distribuzione di probabilità sulle 21 classi e assegna la classe a massima probabilità.

4.4 Procedura di addestramento e valutazione

Il corpus è stato partizionato in un training set (80%) e in un test set (20%) mediante train_test_split con stratificazione sulle etichette (stratify=y) e seed deterministico (random_state=42). La stratificazione garantisce che la proporzione delle classi sia preservata in entrambi i sottoinsiemi, mitigando l'impatto dello sbilanciamento della distribuzione. Il modello è stato addestrato sul training set e valutato sul test set tramite il calcolo dell'accuratezza (accuracy).

4.5 Serializzazione e distribuzione

Il modello addestrato e il vettorizzatore sono serializzati in formato .joblib, scelto per leggerezza e praticità in ambiente Python, e distribuiti congiuntamente: il classificatore e il vettorizzatore devono essere impiegati insieme, poiché il modello opera esclusivamente su rappresentazioni numeriche. Il formato .joblib non consente di recuperare i dati di addestramento originali, garantendo che i testi non siano estraibili dai file distribuiti. Il progetto è reso disponibile gratuitamente e open source su GitHub.

5. Valutazione sperimentale

5.1 Setup sperimentale

La valutazione è condotta sul test set stratificato, composto da 8.935 prompt (20% del corpus), non osservati durante l'addestramento. La metrica primaria adottata è l'accuratezza, definita come il rapporto tra il numero di predizioni corrette e il numero totale di esempi valutati.

5.2 Risultati

Sul test set, il modello raggiunge un'accuratezza complessiva del 81,31% (0,8131). Tale risultato, ottenuto su un problema di classificazione a 21 classi con una distribuzione fortemente sbilanciata, evidenzia la capacità del modello di generalizzare su prompt reali e non osservati, pur con margini di miglioramento nelle classi a bassa cardinalità.

5.3 Discussione e limitazioni

Come qualsiasi sistema di intelligenza artificiale, Emma Guardrails 1.0 non è infallibile e non deve essere considerato una misura di sicurezza esaustiva. Tra le limitazioni si segnalano: (i) lo sbilanciamento delle classi, che penalizza l'apprendimento delle categorie meno rappresentate; (ii) l'impiego di un modello lessicale basato su TF-IDF, che non cattura pienamente fenomeni semantici complessi o dipendenze di lungo raggio; (iii) la natura della metrica di accuratezza, che su dataset sbilanciati può sovrastimare la qualità percepita del classificatore. Inoltre, poiché il vettorizzatore è stato adattato sull'intero corpus prima della partizione, la stima di accuratezza potrebbe risultare leggermente ottimistica; le versioni successive adotteranno un adattamento del vettorizzatore limitato al solo training fold. Ciononostante, il modello costituisce un'efficace prima linea di difesa contro contenuti nocivi, tentativi di abuso e richieste a basso valore aggiunto.

6. Applicazioni

Emma Guardrails può essere integrato in diverse architetture applicative, tra cui:

- Guardrail a monte di LLM: classificazione dei prompt in ingresso e blocco o instradamento delle richieste non conformi prima dell'elaborazione da parte del modello principale.
- Moderazione dell'output: analisi delle risposte generate e blocco preventivo dei contenuti non conformi prima della restituzione all'utente.
- Assistenti virtuali e chatbot aziendali: filtro di sicurezza per applicazioni customer-facing in settori regolamentati.
- Piattaforme educative e per minori: intercettazione di contenuti inappropriati e protezione degli utenti vulnerabili.
- Sistemi di compliance e auditing: rilevazione automatica di richieste illecite o non conformi a supporto dei processi di governance.
- Ricerca accademica: benchmark e baseline per lo sviluppo di modelli di moderazione in lingua italiana.

7. Benefici per il mercato italiano

Il rilascio di Emma Guardrails produce benefici specifici per il contesto italiano. In primo luogo, fornisce uno strumento nativamente italiano, addestrato su prompt reali in lingua italiana, colmando un divario rispetto ai modelli di moderazione prevalentemente anglocentrici. In secondo luogo, la distribuzione open source e gratuita riduce le barriere all'adozione per piccole e medie imprese, amministrazioni pubbliche e realtà del terzo

settore, che possono integrare un livello di protezione senza sostenere costi di licenza. In terzo luogo, il modello contribuisce all'allineamento ai requisiti normativi europei — tra cui il Regolamento (UE) 2024/1689 (AI Act) [9] e il GDPR — offrendo un meccanismo di mitigazione del rischio documentabile e trasparente. Infine, l'efficienza computazionale del modello consente un'esecuzione a bassa latenza e con ridotto consumo di risorse, favorendo il deployment anche in scenari on-premise e con vincoli di costo.

8. Considerazioni etiche, privacy e sicurezza

La progettazione di Emma Guardrails 1.0 è ispirata ai principi di trasparenza, responsabilità e minimizzazione del dato. I prompt utilizzati per l'addestramento sono stati anonimizzati in modo irreversibile e il dataset distribuito non consente la re-identificazione degli utenti. Il modello è concepito come componente di difesa e non come strumento di sorveglianza: la classificazione avviene sul contenuto del singolo prompt e non sull'identità dell'utente. Si evidenzia, inoltre, che il modello non è infallibile e che il suo impiego deve avvenire all'interno di una strategia di sicurezza più ampia, che includa revisione umana, monitoraggio continuo e aggiornamento periodico della tassonomia e dei dati.

9. Conclusioni e sviluppi futuri

Emma Guardrails 1.0 rappresenta un contributo concreto e open source alla sicurezza dei sistemi di intelligenza artificiale in lingua italiana. Con un corpus di 44.677 prompt reali anonimizzati, una tassonomia di 21 categorie e un'architettura TF-IDF + regressione logistica, il modello raggiunge un'accuratezza del 81,31% e si configura come un efficace guardrail a monte per LLM, con ricadute positive in termini di sicurezza, contenimento dei costi e conformità normativa.

Tra gli sviluppi futuri si prevedono: l'ampliamento e il bilanciamento del corpus, l'adozione di rappresentazioni semantiche contestuali (ad esempio modelli transformer pre-addestrati per l'italiano [4], [5]), l'introduzione di metriche di valutazione più robuste per dataset sbilanciati (precision, recall, F1 macro e micro), la calibrazione delle soglie di confidenza e l'estensione del formato di distribuzione verso ONNX per una maggiore portabilità cross-platform.

Riferimenti bibliografici

- [1] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [2] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.
- [3] Crammer, K., & Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2, 265–292.
- [4] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

- [5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019.
- [6] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. Findings of EMNLP 2020.
- [7] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Balle, B., Kasirzadeh, A., Kenton, Z., et al. (2022). Taxonomy of Risks Posed by Language Models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), 214–229.
- [8] Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio (GDPR), Gazzetta ufficiale dell'Unione europea, L 119, 4.5.2016.
- [9] Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio (Artificial Intelligence Act), Gazzetta ufficiale dell'Unione europea, L, 12.7.2024.