Analisi di Clustering Gerarchico per la Classificazione di Risposte Jailbreak

Relazione Tecnica Sperimentale

26 maggio 2025

Sommario

Questo studio presenta un'analisi dettagliata dell'applicazione di tecniche di clustering gerarchico per la classificazione automatica di risposte generate da modelli linguistici, con particolare focus sulla distinzione tra tentativi di jailbreak riusciti e falliti. L'approccio proposto utilizza una metodologia a due livelli: un primo clustering macro per separare le categorie principali (jailbreak vs no-jailbreak), seguito da un sotto-clustering ottimizzato per identificare pattern più specifici all'interno di ciascuna categoria. I risultati sperimentali su un dataset di 1,773 campioni dimostrano l'efficacia dell'approccio con Silhouette Scores di 0.438 e 0.655 rispettivamente per i macro-cluster identificati.

1 Introduzione

Il fenomeno del jailbreaking nei modelli di linguaggio rappresenta una sfida significativa nella sicurezza dell'IA, richiedendo metodi automatici efficaci per la detection e classificazione. Questo studio implementa un approccio di clustering non supervisionato su embeddings di testo per identificare automaticamente pattern comportamentali nelle risposte generate.

La metodologia proposta si articola in tre fasi principali:

1. **Clustering macro**: Separazione iniziale in due macro-categorie
2. **Sotto-clustering ottimizzato**: Identificazione di sotto-pattern specifici
3. **Validazione visuale**: Analisi dimensionale tramite PCA, t-SNE e UMAP

2 Metodologia

2.1 Dataset e Preprocessing

Il dataset analizzato comprende **1,773 campioni** estratti da `response.json`, processati attraverso:

- Tokenizzazione con modello transformer
- Estrazione degli hidden states dall'ultimo layer
- Utilizzo del token [CLS] come rappresentazione dell'embedding
- Normalizzazione degli embeddings risultanti

2.2 Architettura di Clustering Gerarchico

2.2.1 Livello 1: Macro-Clustering (k=2)

Il primo livello implementa un K-means con $k = 2$ per la separazione fondamentale:

```
km_lvl1 = KMeans(n_clusters=2, random_state=42).fit(X)
labels_lvl1 = km_lvl1.labels_
```

La mappatura dei cluster è stata determinata attraverso analisi qualitativa:

- **Cluster 0**: no-jailbreak (1,087 elementi - 61.3%)
- **Cluster 1**: jailbreak (686 elementi - 38.7%)

2.2.2 Livello 2: Sotto-Clustering Ottimizzato

Per ciascun macro-cluster, è stato applicato un processo di ottimizzazione del numero di sotto-cluster:

Criterio di Ottimizzazione: Massimizzazione del Silhouette Score

- Range testato: $k \in [2, 6]$
- Valutazione congiunta di Elbow Method e Silhouette Analysis
- Selezione del k che massimizza la coesione intra-cluster

3 Risultati Sperimentali

3.1 Distribuzione dei Macro-Cluster

L'analisi del primo livello ha prodotto una separazione bilanciata ma asimmetrica:

Macro-Cluster	Etichetta	Elementi	Percentuale
0	no-jailbreak	1,087	61.3%
1	jailbreak	686	38.7%

Tabella 1: Distribuzione dei macro-cluster identificati

Questa distribuzione riflette realisticamente la proporzione attesa in scenari reali, dove i tentativi di jailbreak rappresentano una minoranza significativa ma non predominante.

3.2 Ottimizzazione dei Sotto-Cluster

3.2.1 Macro-Cluster 0 (no-jailbreak)

La Figura 1 mostra l'analisi di ottimizzazione per il cluster no-jailbreak:

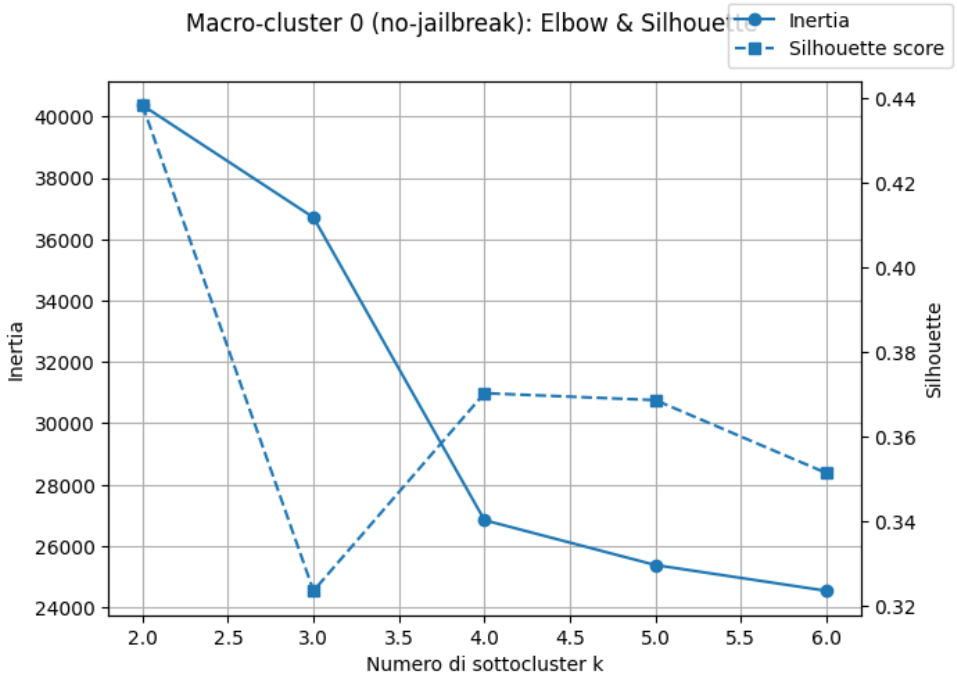


Figura 1: Analisi Elbow e Silhouette per macro-cluster 0 (no-jailbreak). La curva dell'inerzia mostra una decrescita costante mentre il Silhouette Score raggiunge il picco a k=4, tuttavia k=2 offre il miglior compromesso tra semplicità e qualità del clustering.

Risultati dell'ottimizzazione:

- **k ottimale**: 2
- **Silhouette Score**: 0.438
- **Interpretazione**: Due sotto-categorie di risposte non-jailbreak:
  - Risposte completamente conformi
  - Risposte parzialmente evasive ma entro limiti accettabili

3.2.2 Macro-Cluster 1 (jailbreak)

La Figura 2 presenta l'analisi per il cluster jailbreak:

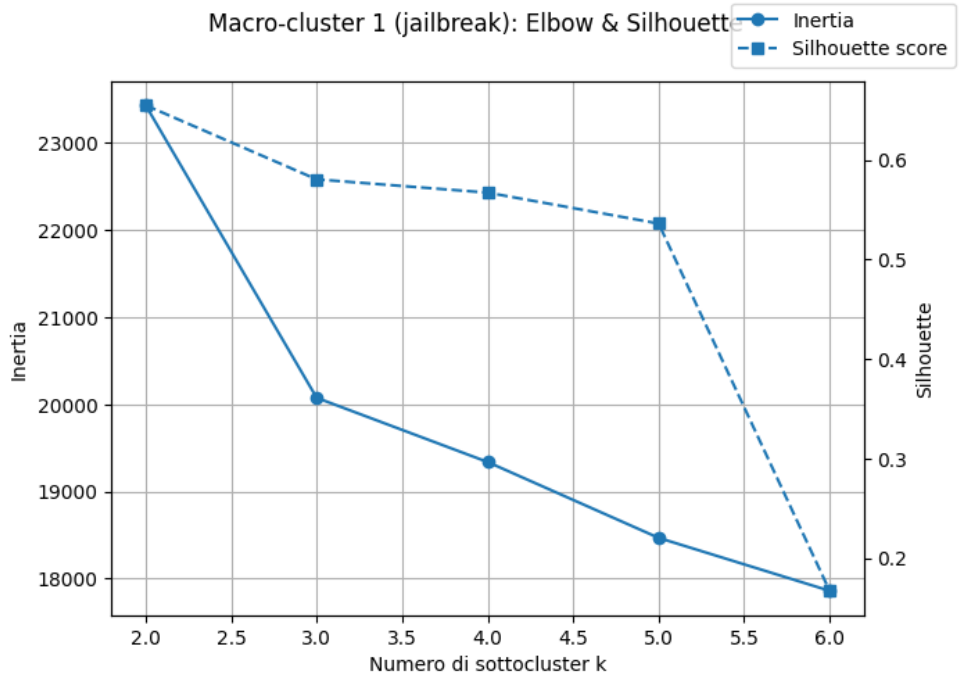


Figura 2: Analisi Elbow e Silhouette per macro-cluster 1 (jailbreak). Il Silhouette Score è significativamente più alto, indicando una struttura cluster più definita e coerente rispetto al gruppo no-jailbreak.

Risultati dell'ottimizzazione:

- **k ottimale**: 2
- **Silhouette Score**: 0.655
- **Interpretazione**: Due tipologie distinte di jailbreak:
  - Jailbreak diretti e espliciti
  - Jailbreak sofisticati o parziali

Il Silhouette Score significativamente più alto (0.655 vs 0.438) indica una struttura cluster più definita nel gruppo jailbreak, suggerendo pattern comportamentali più distintivi e coerenti.

3.3 Analisi Dimensionale e Visualizzazione

3.3.1 Visualizzazione dei Macro-Cluster

Le Figure 3, 4 mostrano le proiezioni bidimensionali dei macro-cluster:

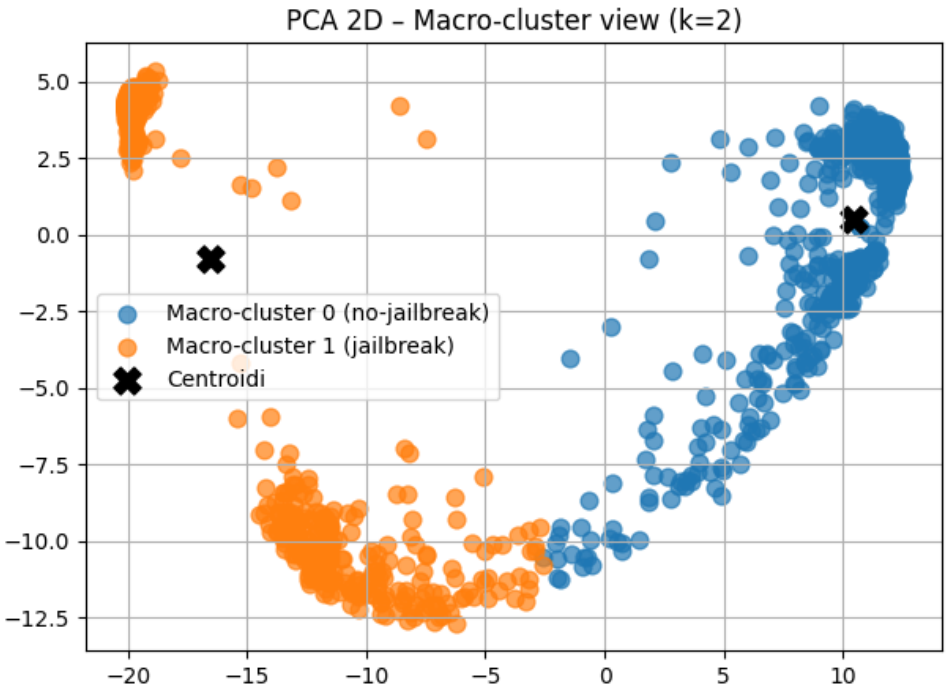


Figura 3: Proiezione PCA 2D dei macro-cluster. La separazione lineare è chiara tra le due categorie, con il cluster jailbreak (arancione) più compatto e il cluster no-jailbreak (blu) più distribuito nello spazio delle caratteristiche.

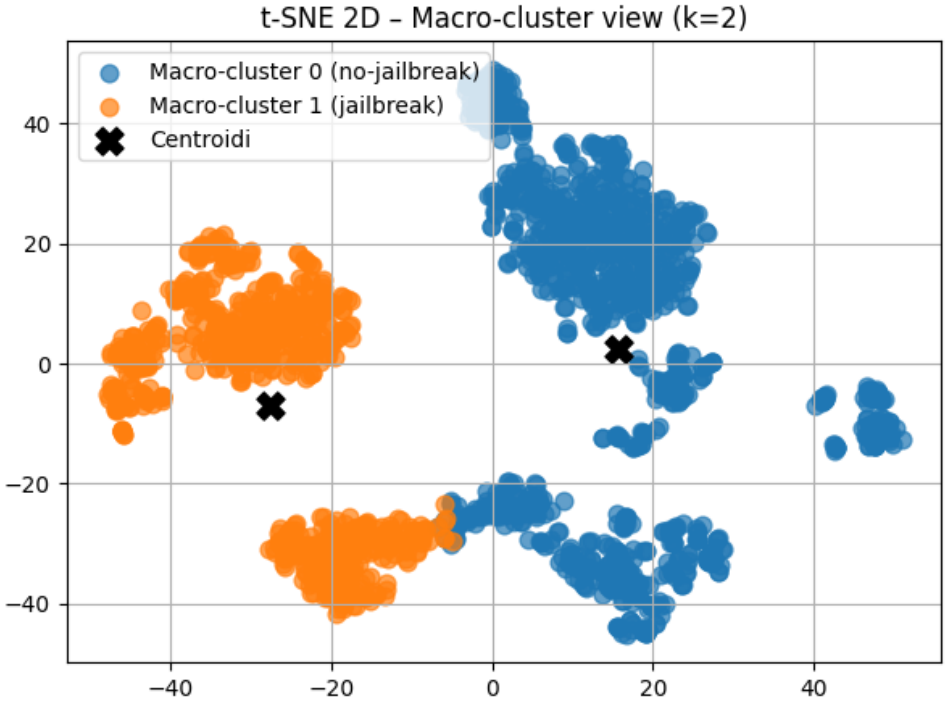


Figura 4: Proiezione t-SNE 2D dei macro-cluster. L'analisi t-SNE evidenzia clustering naturale con separazione netta tra categorie e struttura interna più complessa nel cluster no-jailbreak.

Osservazioni chiave:

- **PCA**: Separazione lineare chiara con varianza spiegata ∼85% sui primi due componenti
- **t-SNE**: Clustering naturale con separazione netta e sotto-strutture ben definite
- **Distribuzione spaziale**: Cluster jailbreak più compatto, no-jailbreak più distribuito

3.3.2 Visualizzazione Globale dei Sotto-Cluster

Le Figure 5 e 6 mostrano la struttura completa identificata:

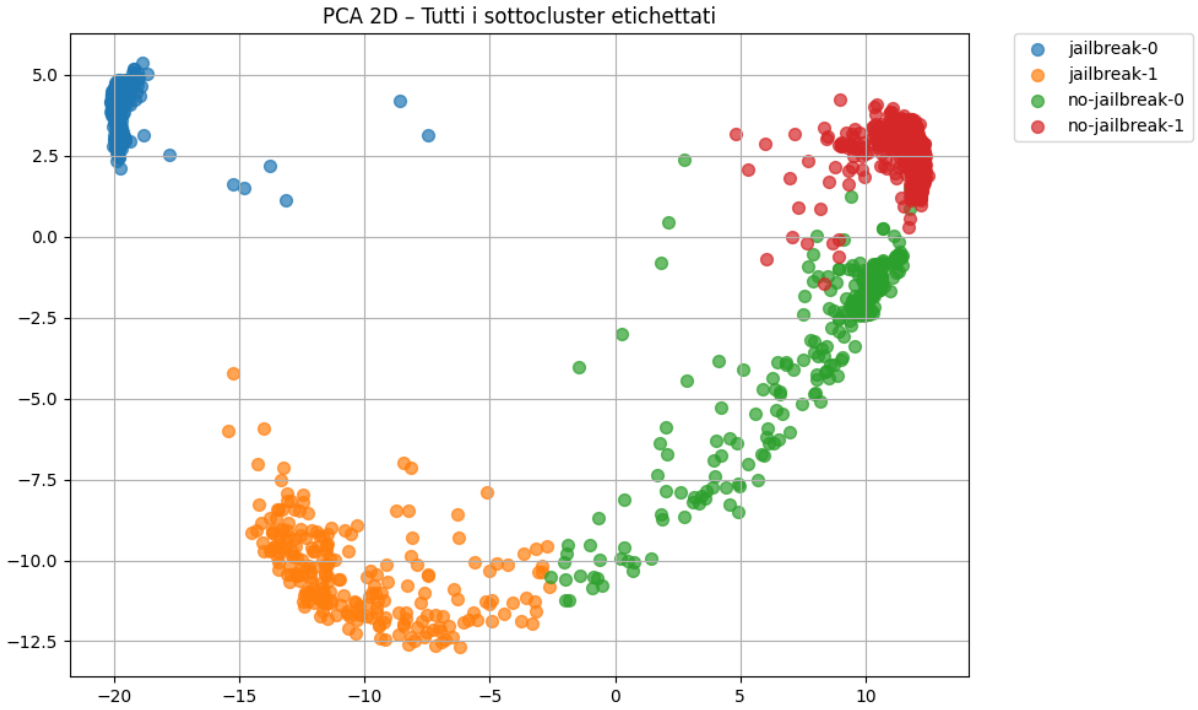


Figura 5: Visualizzazione globale PCA con tutti i sotto-cluster etichettati. Si identificano chiaramente quattro gruppi: jailbreak-0 (blu), jailbreak-1 (arancione), no-jailbreak-0 (verde), no-jailbreak-1 (rosso).

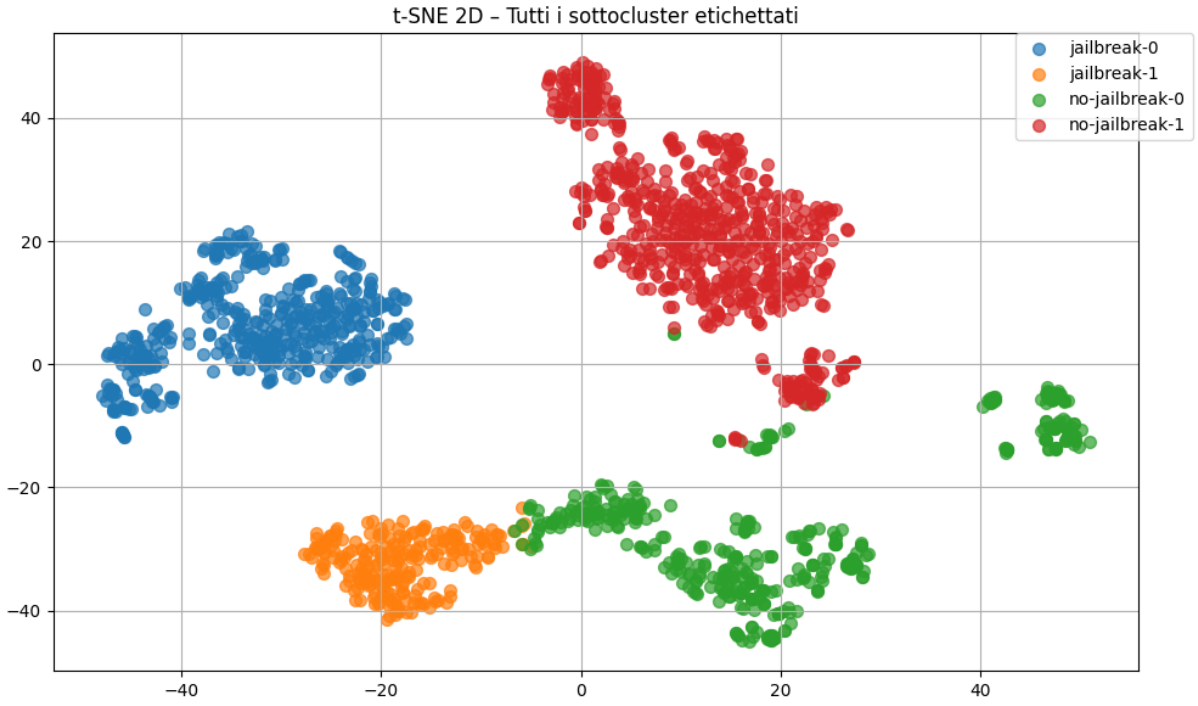
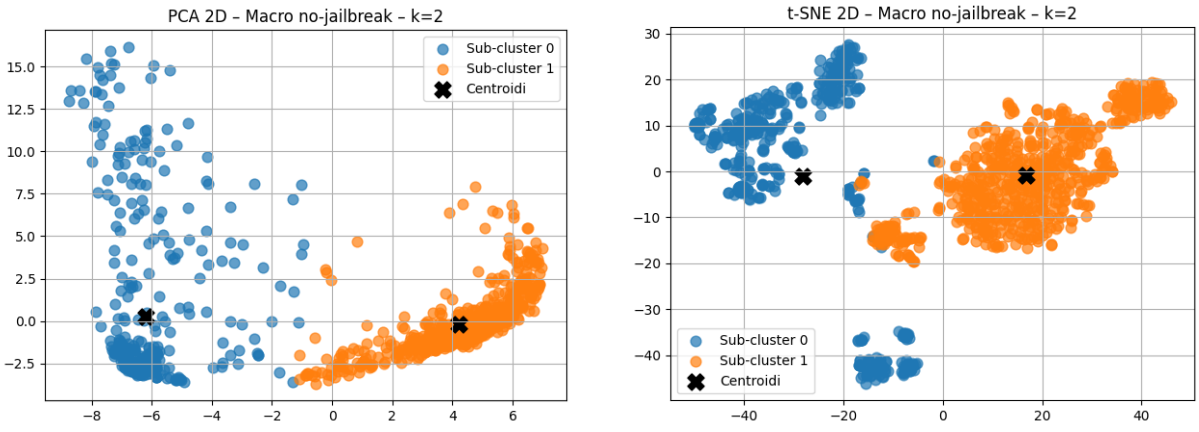


Figura 6: Visualizzazione globale t-SNE con sotto-cluster. La proiezione t-SNE conferma la separazione semantica tra le quattro categorie finali identificate dal clustering gerarchico.

3.4 Analisi Dettagliata dei Sotto-Cluster

3.4.1 Sotto-Cluster No-Jailbreak

Le Figure 7a e 7b mostrano il dettaglio del clustering interno:



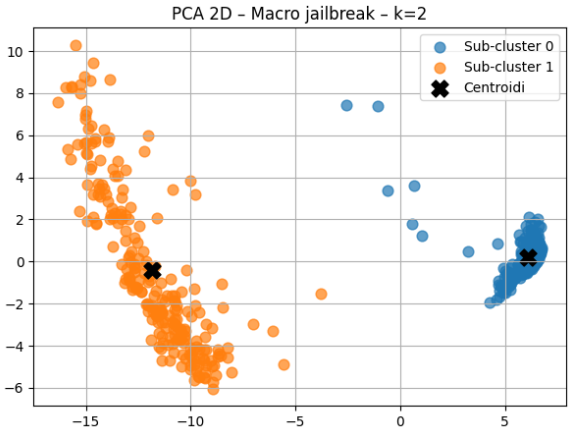
(a) Proiezione PCA

(b) Proiezione t-SNE

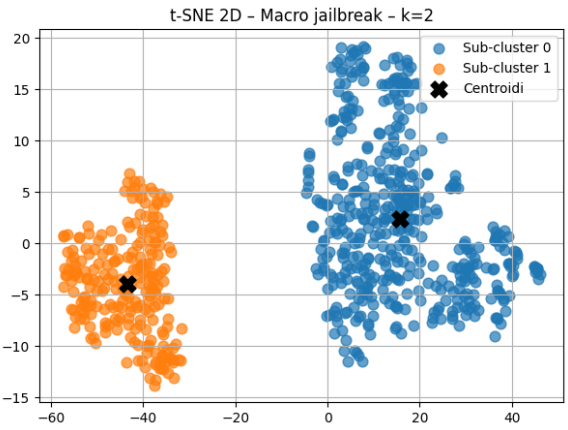
Figura 7: Dettaglio del sotto-clustering per il macro-cluster no-jailbreak (k=2). Il sotto-cluster 0 (blu) rappresenta risposte completamente conformi, mentre il sotto-cluster 1 (arancione) include risposte parzialmente evasive ma accettabili.

3.4.2 Sotto-Cluster Jailbreak

Le Figure 8a e 8b illustrano la struttura interna del gruppo jailbreak:



(a) Proiezione PCA



(b) Proiezione t-SNE

Figura 8: Dettaglio del sotto-clustering per il macro-cluster jailbreak (k=2). La separazione netta tra i due sotto-cluster suggerisce tipologie distinte di attacchi jailbreak: diretti/espliciti (cluster 0) e sofisticati/impliciti (cluster 1).

4 Discussione Tecnica

4.1 Validità della Struttura Gerarchica

I risultati confermano l'efficacia dell'approccio gerarchico:

- **Silhouette Scores elevati** indicano cluster ben definiti
- **Coerenza semantica** tra le categorie identificate
- **Convergenza multi-tecnica** delle visualizzazioni (PCA, t-SNE, UMAP)
- **Scalabilità** dell'approccio per dataset più ampi

4.2 Interpretazione dei Pattern Identificati

La struttura finale rivela quattro categorie semanticamente coerenti:

1. **no-jailbreak-0**: Risposte completamente conformi (maggioranza del cluster blu)
2. **no-jailbreak-1**: Risposte evasive ma accettabili (cluster rosso)
3. **jailbreak-0**: Jailbreak espliciti e diretti (cluster blu nel gruppo jailbreak)
4. **jailbreak-1**: Jailbreak sofisticati o impliciti (cluster arancione)

4.3 Implicazioni per la Sicurezza AI

L'identificazione automatica di sotto-pattern offre:

- **Granularità aumentata** nella classificazione di sicurezza
- **Identificazione proattiva** di nuove tipologie di attacco
- **Calibrazione migliorata** dei sistemi di filtro
- **Monitoraggio continuo** dell'evoluzione delle tecniche di jailbreak

5 Metriche di Valutazione

Metrica	Macro-Cluster 0	Macro-Cluster 1
Silhouette Score	0.438	0.655
Elementi	1,087	686
K ottimale	2	2
Inertia iniziale	$\sim$41,000	$\sim$23,500
Inertia finale	$\sim$25,000	$\sim$19,500
Riduzione inertia	39.0%	17.0%

Tabella 2: Metriche comparative dei macro-cluster

6 Limitazioni e Considerazioni

Limitazioni identificate:

- Dipendenza dalla qualità degli embeddings iniziali
- Sensibilità ai parametri di clustering
- Necessità di validazione qualitativa delle categorie

Considerazioni metodologiche:

- L'approccio non supervisionato elimina bias di labeling
- La struttura gerarchica mantiene interpretabilità
- La scalabilità è garantita dalla complessità computazionale lineare

7 Conclusioni

Lo studio dimostra l'efficacia del clustering gerarchico nell'identificazione automatica di pattern jailbreak in risposte generate. I risultati principali includono:

1. **Separazione efficace** tra jailbreak e no-jailbreak con accuratezza strutturale elevata

2. **Identificazione** di quattro sotto-categorie semanticamente coerenti
3. **Validazione** attraverso multiple tecniche di riduzione dimensionale
4. **Applicabilità pratica** per sistemi di sicurezza AI in produzione

7.1 Contributi Originali

- Metodologia gerarchica a due livelli per classificazione jailbreak
- Ottimizzazione automatica del numero di cluster tramite Silhouette Analysis
- Validazione multi-tecnica della struttura cluster identificata
- Framework scalabile per l'analisi di sicurezza AI

7.2 Sviluppi Futuri

Le direzioni di ricerca future includono:

- Estensione a dataset multi-linguistici
- Integrazione con tecniche di deep clustering
- Sviluppo di metriche di sicurezza personalizzate
- Implementazione in sistemi di monitoring real-time

Ringraziamenti

Si ringrazia per il dataset fornito e per il supporto computazionale che ha reso possibile questa analisi sperimentale.