

# Prober Universale

Rilevamento delle allucinazioni nei Large Language Models

Emanuele Fontana

Università degli Studi di Bari Aldo Moro  
Dipartimento di Informatica

# Introduzione

---

- L'avvento di GPT-3 e degli LLM ha rivoluzionato il NLP.
- Gli LLM sono strumenti potenti ma soffrono di **allucinazioni**.
- **Allucinazione:** generazione di testo sintatticamente corretto ma fattualmente errato.
- Esempio:

*Utente: How many r's are there in the word strawberries?*

*LLM: 2*

# Introduzione

---

- Le allucinazioni limitano l'uso degli LLM in ambiti critici (medicina, legge).
- Tecniche attuali (RAG, CoT) non sono infallibili.
- Necessità di rilevare quando un modello sta allucinando.

# Introduzione e obiettivi

- **Probing:** analisi delle attivazioni interne del modello.
- **Ipotesi:** la veridicità è codificata nello spazio latente.

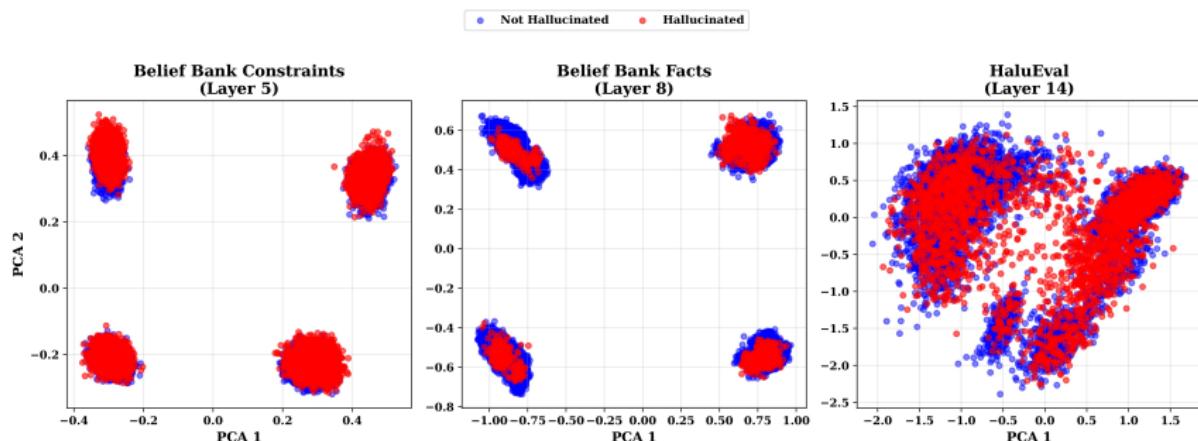


Figura: Esempio sul layer attn di LLaMA3.1-8B-Instruct

- **Obiettivo:** creare un *prober universale*.

# Metodologia: Dataset

---

- **Belief Bank Facts:**
  - Fatti affermativi e negati (es. "An eagle is a bird").
  - 27.416 affermazioni (Bilanciato).
  - Rilevamento *Factual Hallucinations*.
- **Belief Bank Constraints:**
  - Implicazioni e mutue esclusioni.
  - 25.756 affermazioni.
  - Rilevamento *Logical Inconsistencies*.
- **HaluEval:**
  - Contesti conversazionali complessi.
  - 10.000 esempi.

# Metodologia: LLM utilizzati

---

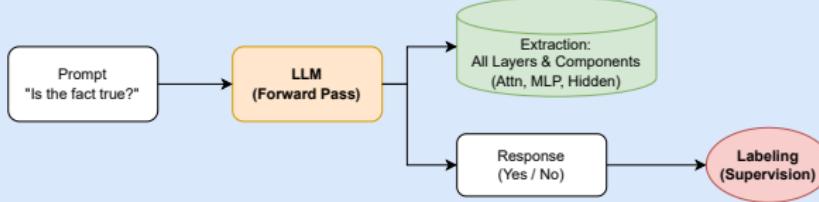
Sono stati utilizzati i seguenti LLM:

- Qwen2.5-7B
- Falcon3-7B-Base
- gemma2-9b-it
- LLama3.1-8B-Instruct

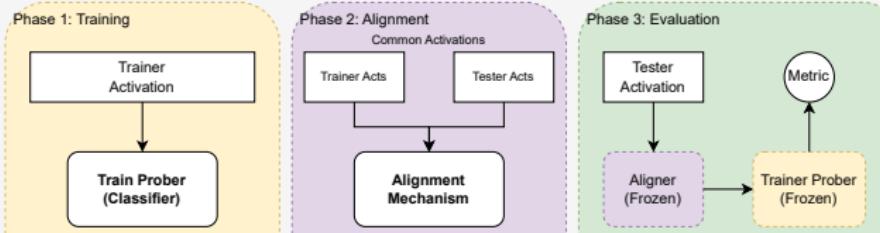
# Metodologia

## Metodologia: Pipeline di Estrazione e Addestramento

### 1. PIPELINE DI ESTRAZIONE



### 2. PIPELINE GENERALE: Addestramento & Allineamento



## Esempio allineamento

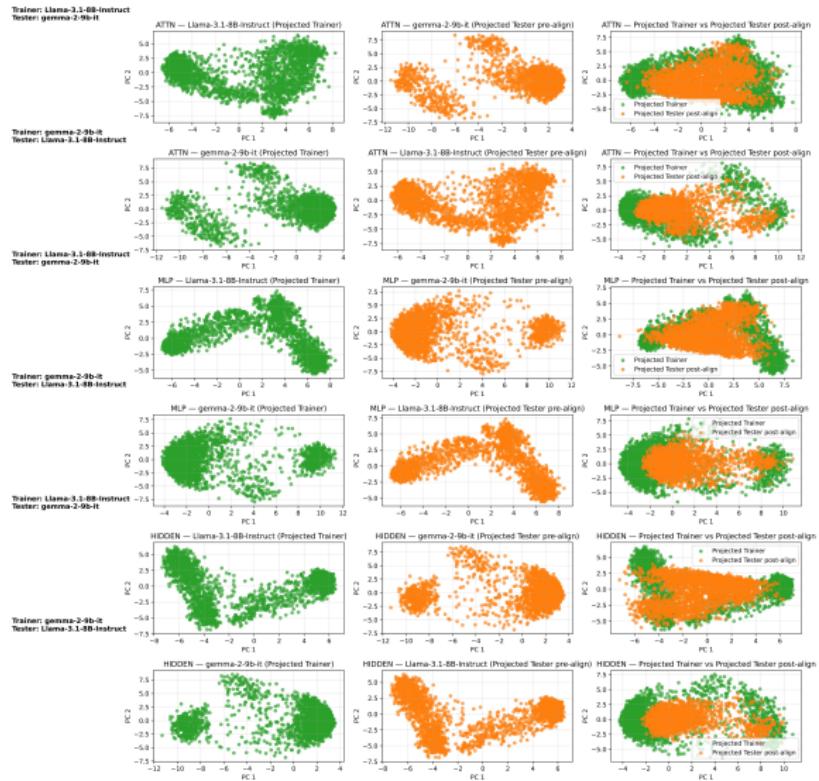


Figura: Esempio allineamento Trainer=LLama, Tester=Gemma

# Metodologia: FullLinear

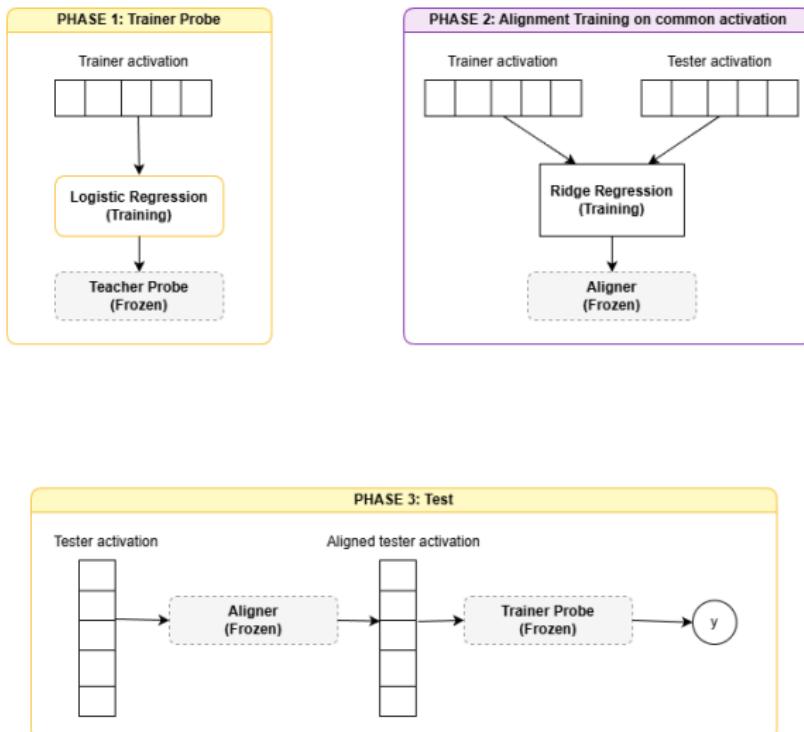


Figura: Pipeline FullLinear: Logistic Regression su Trainer, Ridge Regression per allineamento Tester.

# Metodologia: Approccio AdapterMLP

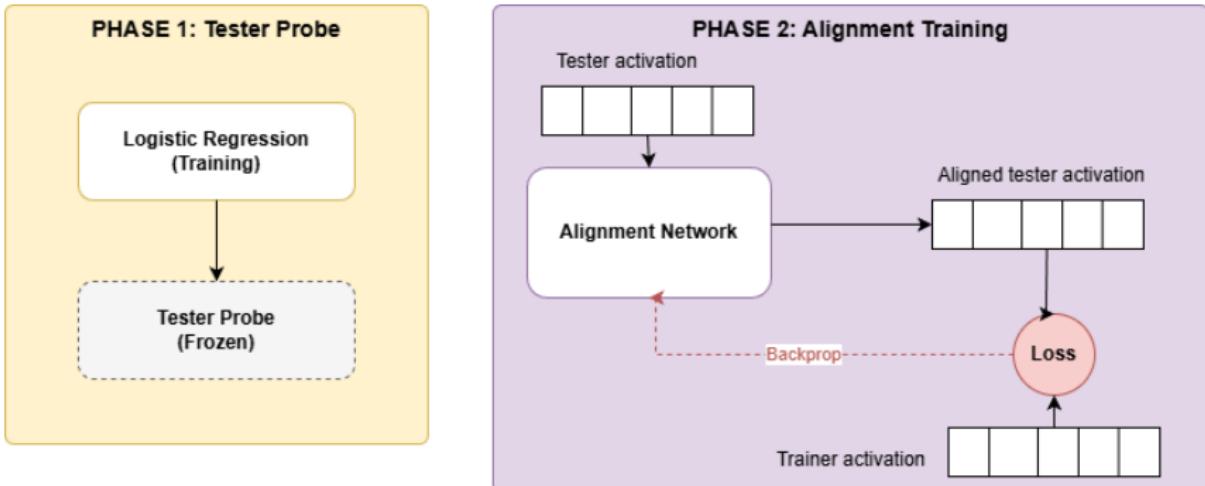


Figura: Pipeline Ibrida: AlignmentNetwork non-lineare per proiettare il Tester, classificatore lineare fisso.

# Metodologia: Approccio Non-Lineare Completo

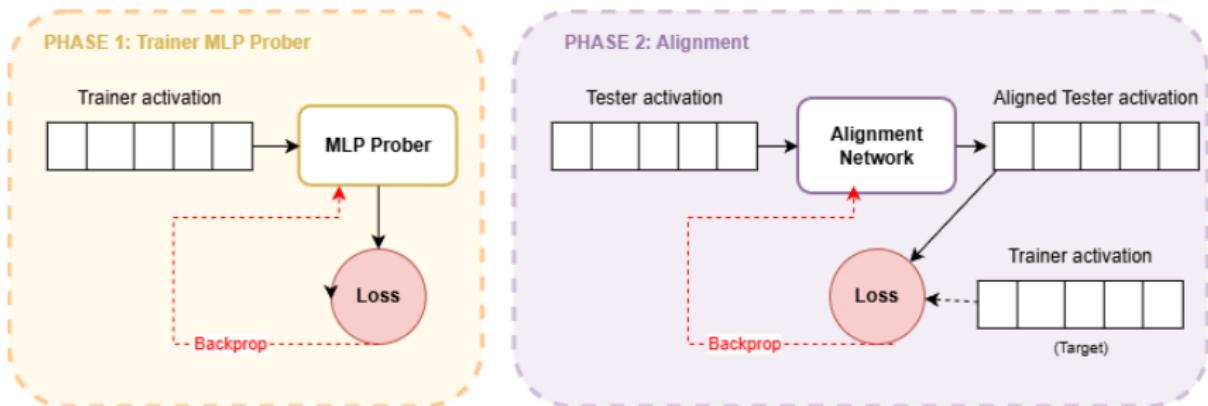


Figura: Pipeline Completa: AlignmentNetwork e MLP Prober entrambi non-lineari.

# Metodologia: Approccio Non-Lineare Ridotto

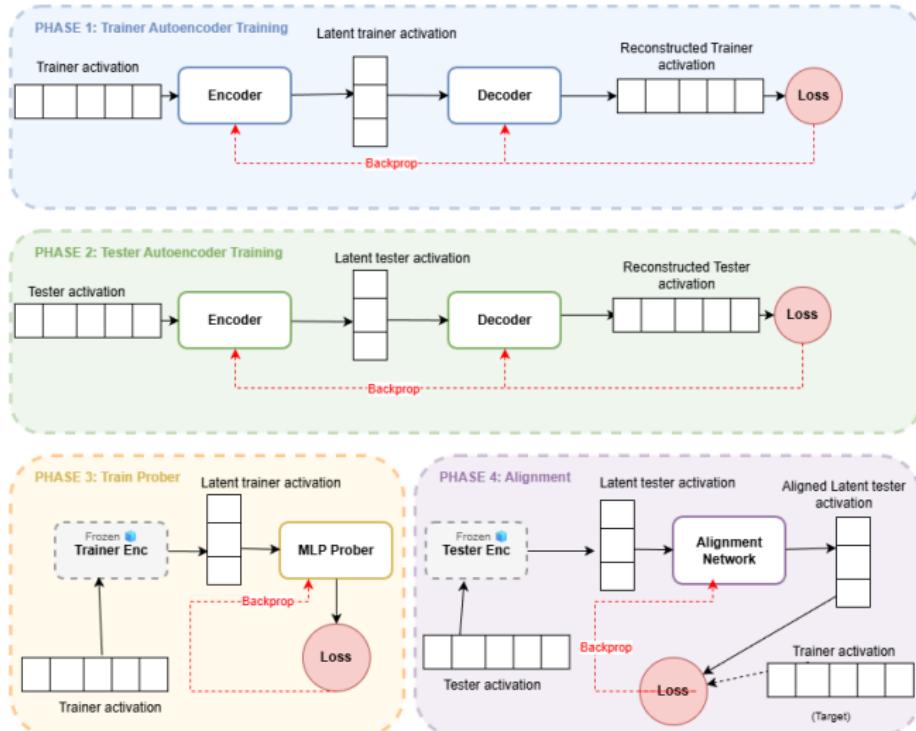


Figura: Pipeline Ridotta: Autoencoder per ridurre il rumore, poi allineamento nello spazio latente.

# Metodologia: One-For-All

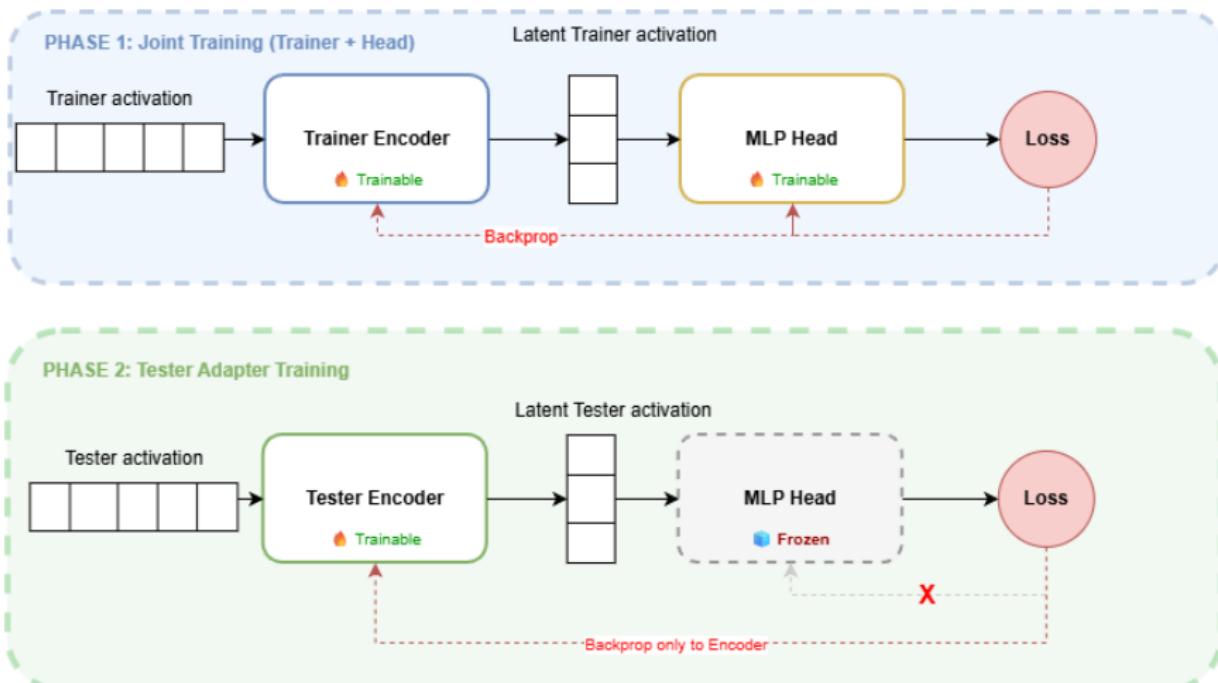


Figura: Pipeline One-For-All: Encoder specifico per modello, Classification Head congelata dal Trainer.

# Results Gemma -> Llama

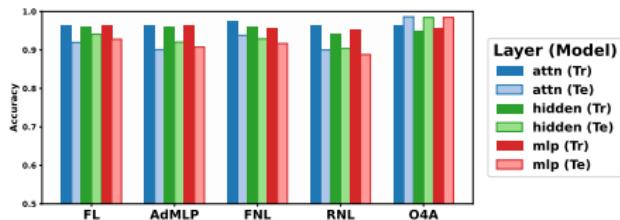


Figura: Factual Hallucinations

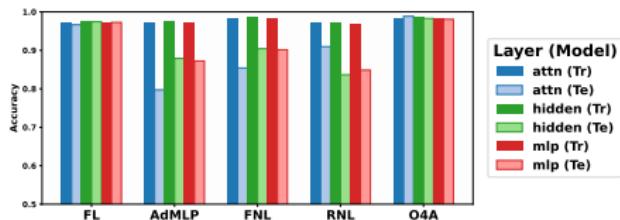


Figura: Logical Hallucinations

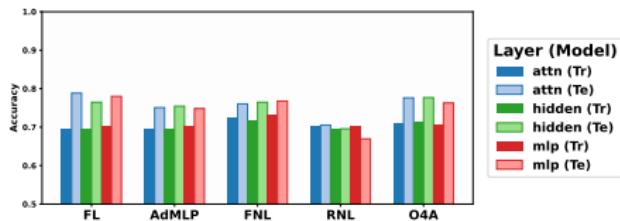


Figura: Contextual Hallucinations

# Cross Domain

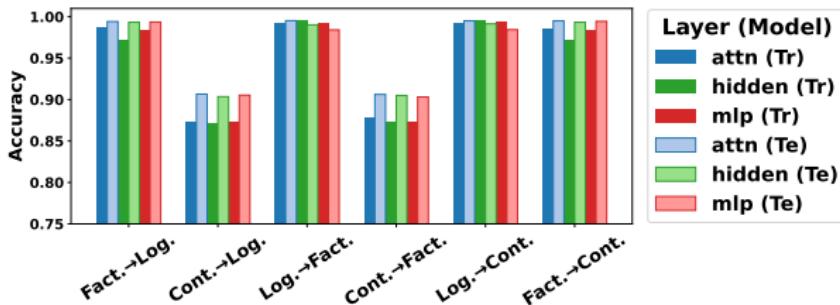


Figura: Trainer:Gemma, Tester:LLama

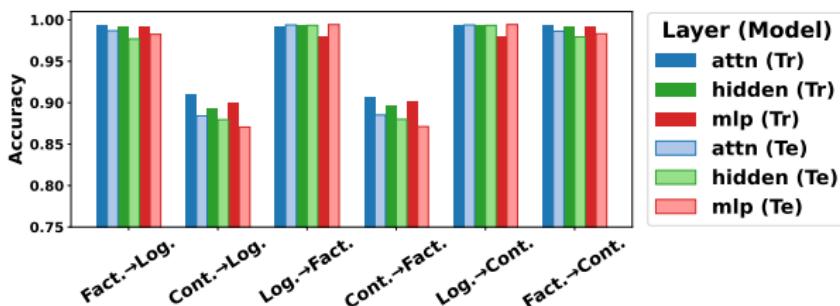


Figura: Trainer:LLama, Tester:Gemma

# Discussion

---

- I metodi di trasferibilità sono generalmente efficaci.
- **One-For-All** emerge come il metodo più promettente e robusto. Questo probabilmente è dovuto all'addestramento End-To-End di Encoder e Classification Head che forza l'encoder a imparare una rappresentazione utile ai fini della classificazione
- La trasferibilità è particolarmente alta per task di verifica fattuale semplice.
  - **Ipotesi:** Le strutture latenti apprese su fatti semplici sono fondamentali e universali. Quelle apprese su contesti complessi sono più rumorose o specifiche.

# Lavori futuri

---

- **Multimodalità:** Estendere l'approccio a modelli Vision-Language.
- **Miglioramento Allineamento:** Esplorare tecniche non lineari più avanzate per task complessi.
- **Studi cross-domain aggiuntivi:** Testare le attivazioni di un dataset sull'encoder di un altro dataset usando la head di classificazione addestrata su un terzo dataset diverso
- **Applicazioni Real-Time:** Implementare il prober in sistemi di monitoraggio live.

Grazie per l'attenzione! 