# Prober Universale

## Rilevamento delle allucinazioni nei Large Language Models

Emanuele Fontana
Università degli Studi di Bari Aldo Moro
Dipartimento di Informatica

23 dicembre 2025

# 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

- **Probing**: analisi delle attivazioni interne del modello.
- Ipotesi: la veridicità è codificata nello spazio latente.
- Obiettivo: creare un **prober universale**.
- Trasferire la capacità di rilevamento da un modello all'altro senza supervisione aggiuntiva.

# Obiettivi del progetto

- Analizzare le attivazioni interne di diverse famiglie di LLM (Qwen, Falcon, Llama, Gemma).
- Identificare pattern comuni associati alle allucinazioni.
- Sviluppare metodologie di allineamento tra spazi latenti.
- Costruire e validare il prober.

# Background

- **Transformer** (Vaswani et al., 2017): Architettura basata su Self-Attention.
- Elaborazione parallela dell'intera sequenza.
- **Self-Attention**:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

- Permette di catturare relazioni a lungo raggio.

# Background

- Modelli probabilistici addestrati su enormi corpora (Next Token Prediction).
- **Scaling Laws**: Prestazioni migliorano con parametri e dati.
- Ciclo di vita:
  1. Pre-training (Unsupervised)
  2. Supervised Fine-Tuning (SFT)
  3. Alignment (RLHF/DPO)

# Background

- **Tipologie di allucinazione**:
  - Fattuali (Falsità oggettive)
  - Logiche (Incoerenze nel ragionamento)
  - Contestuali (Contraddizione del prompt)
- **Cause**: Compressione con perdita, dati rumorosi, Sycophancy.

# Metodologia

- **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

1. Prompting del modello (es. "Is the fact true?").
2. Estrazione attivazioni da tutti i layer e componenti (Attention, MLP, Hidden).
3. Etichettatura basata sulla risposta generata (Yes/No).

# Metodologia: Baseline

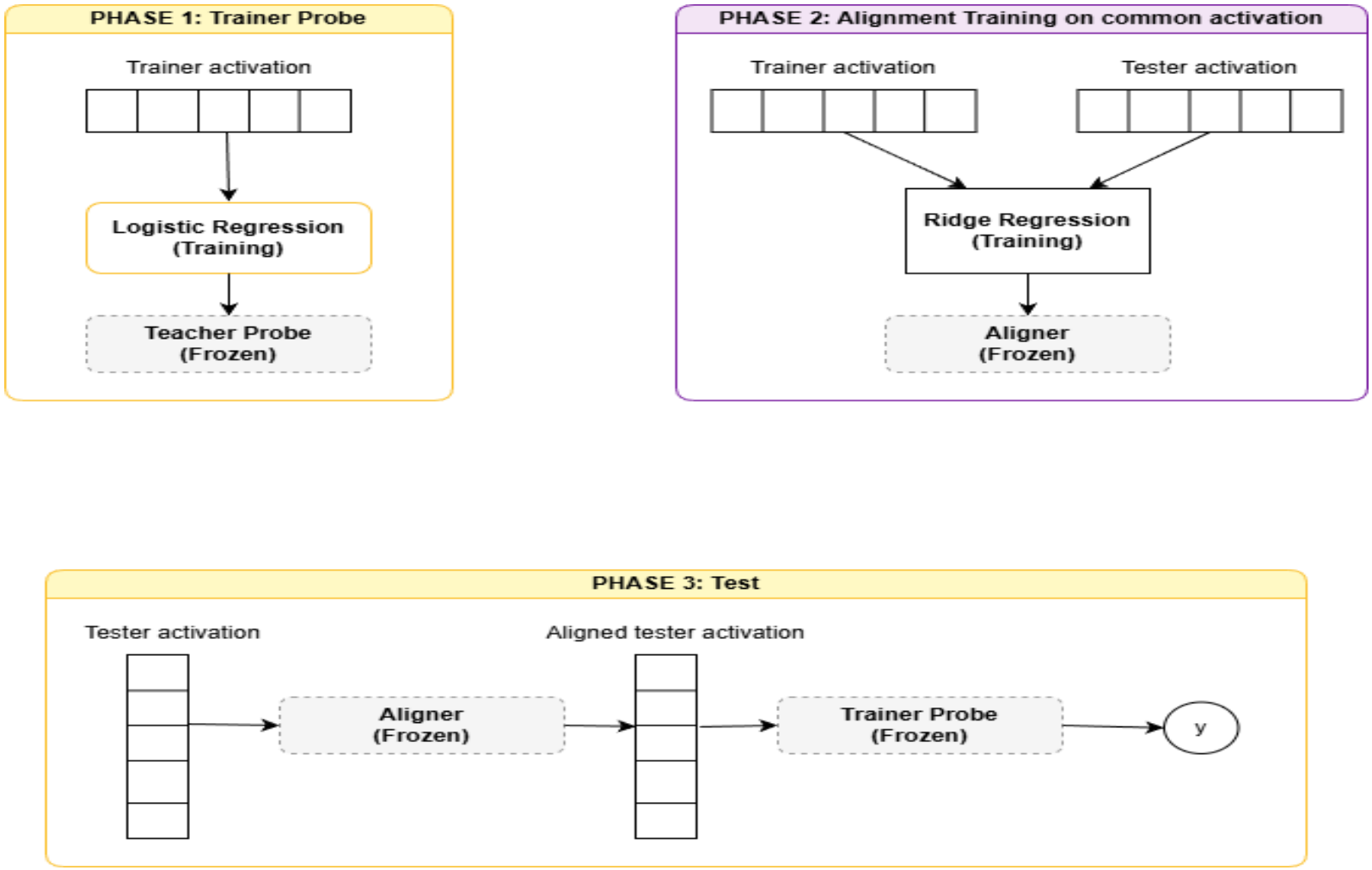


Figura: Pipeline Baseline: Logistic Regression su Trainer, Ridge Regression per allineamento Tester.

# Metodologia: Approccio Ibrido

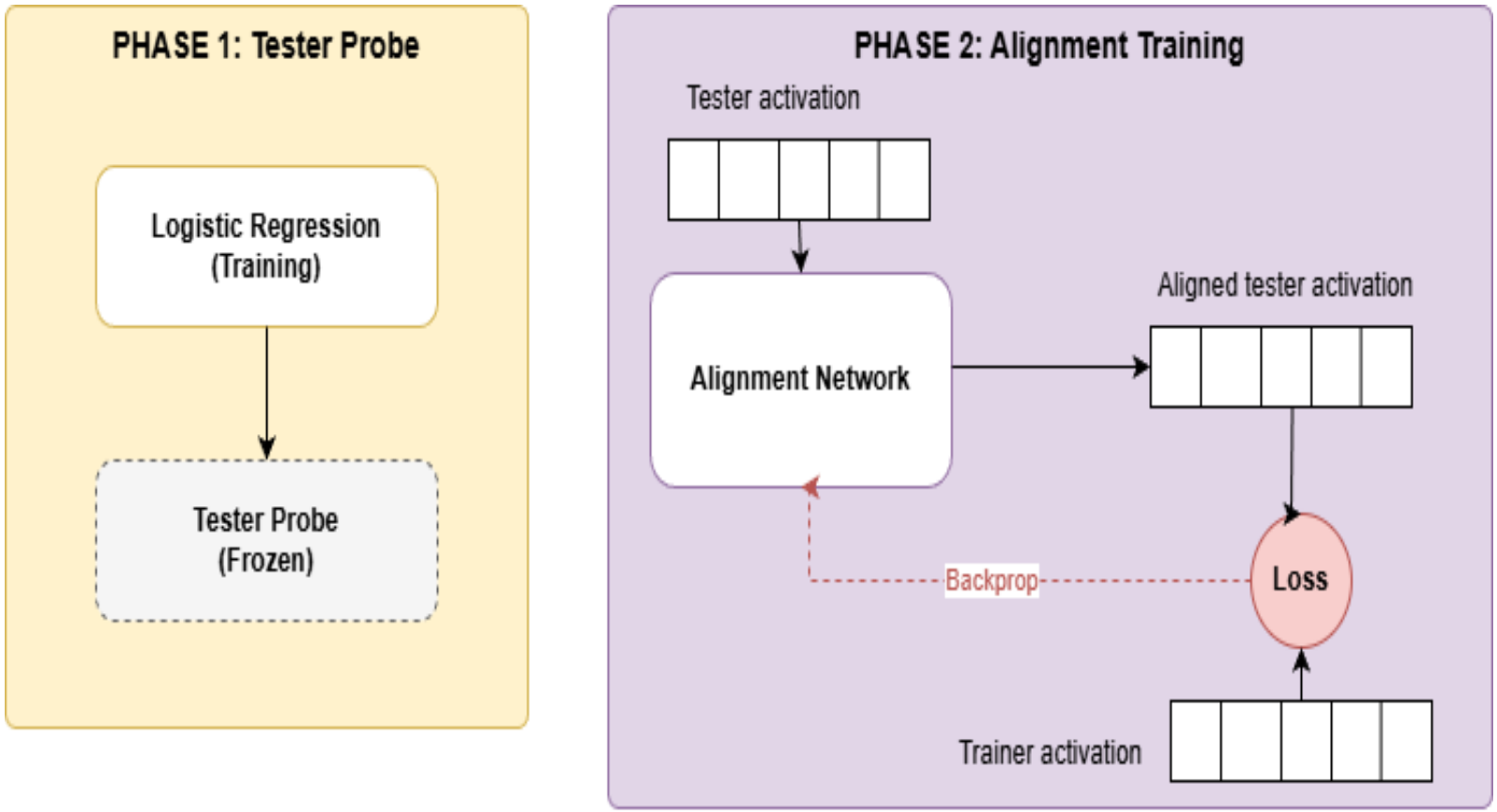


Figura: Pipeline Ibrida: AlignmentNetwork non-lineare per proiettare il Tester, classificatore lineare fisso.

# Metodologia: Approccio Non-Lineare Completo

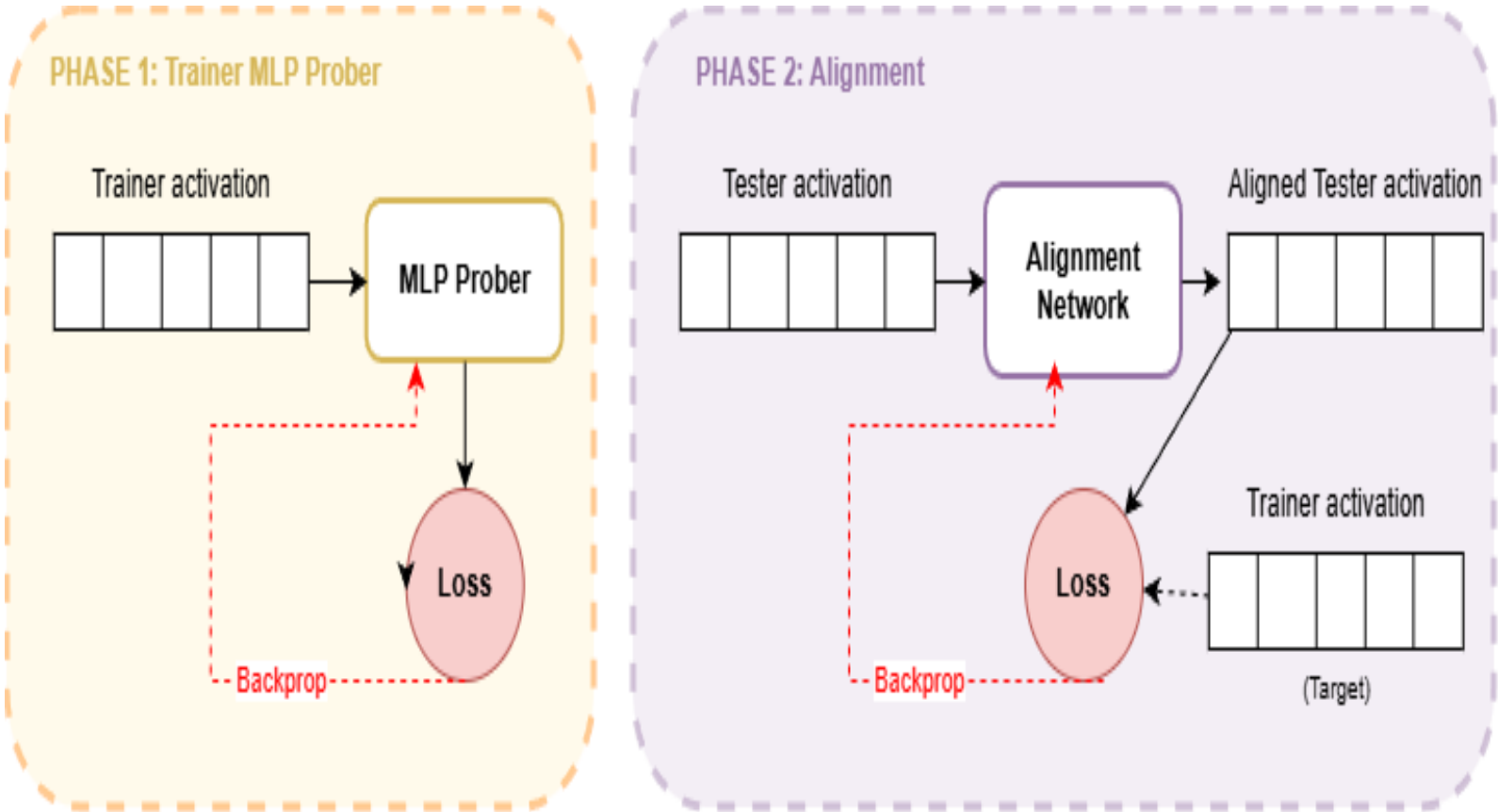


Figura: Pipeline Completa: AlignmentNetwork e MLP Prober entrambi non-lineari.

# Metodologia: Approccio Non-Lineare Ridotto

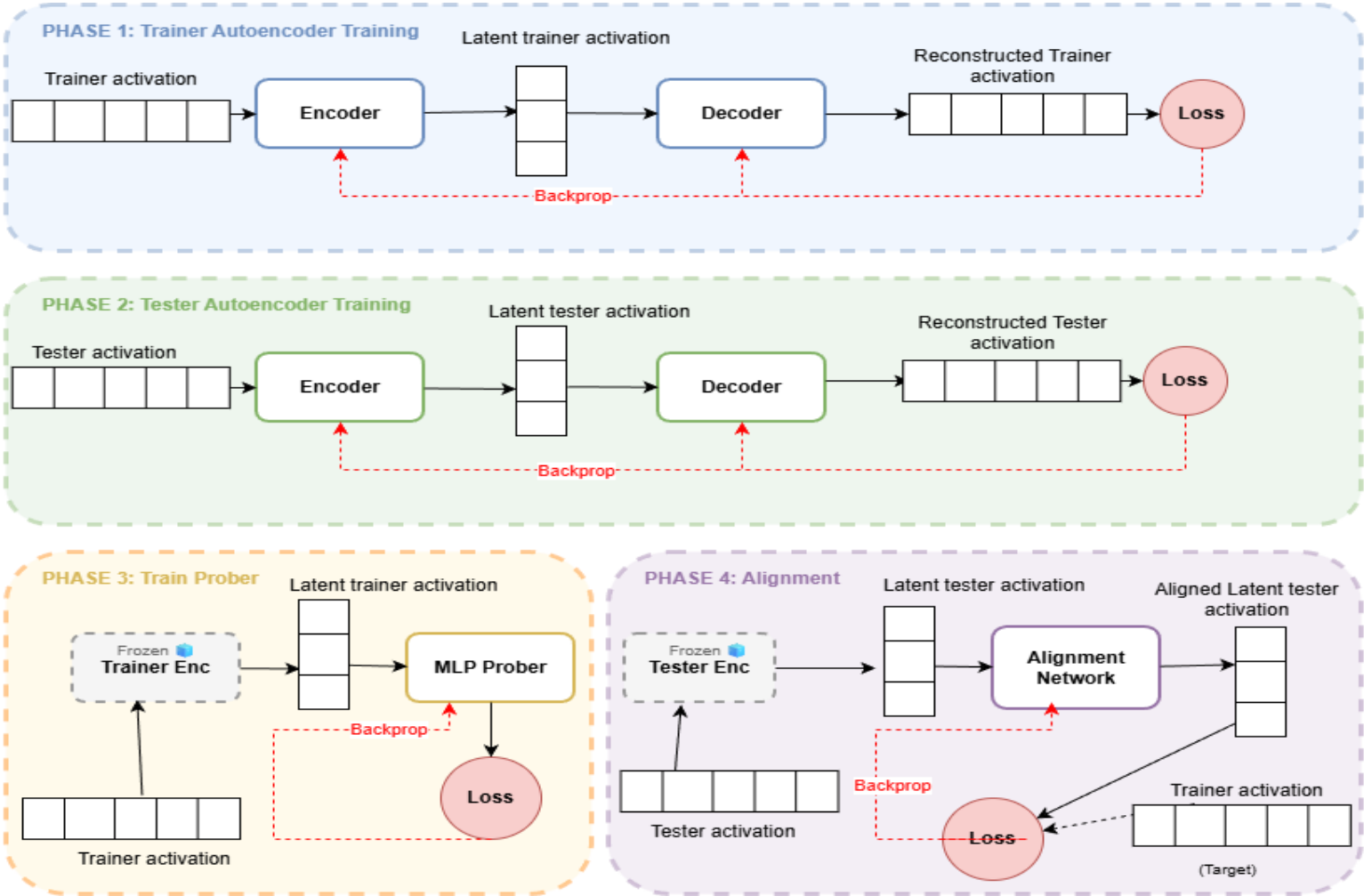


Figura: Pipeline Ridotta: Autoencoder per ridurre il rumore, poi allineamento nello spazio latente.

# Metodologia: One-For-All

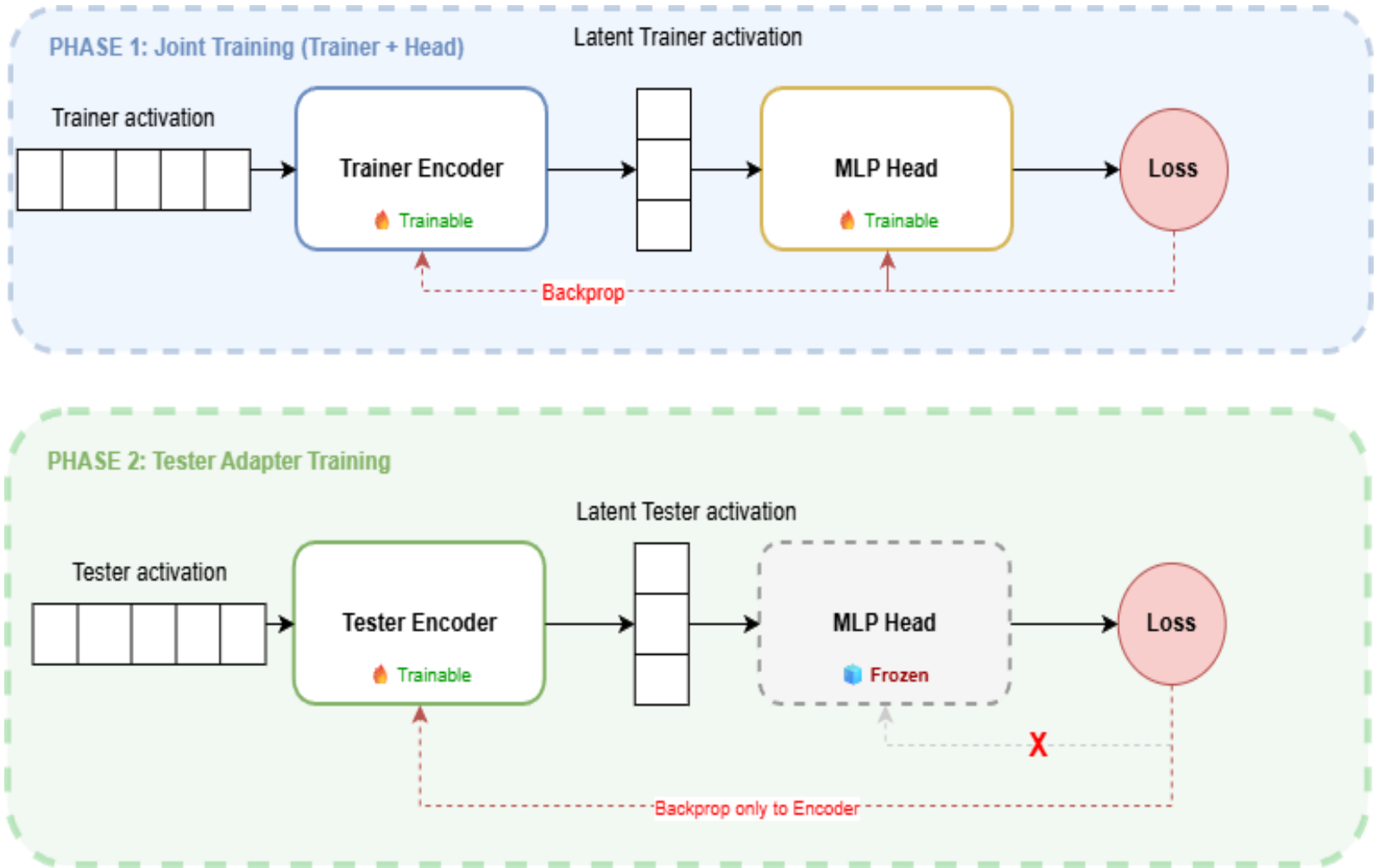


Figura: Pipeline One-For-All: Encoder specifico per modello, Classification Head congelata dal Trainer.

# Metodologia

| Modello | Allucinazioni | % |
|---|---|---|
| Qwen2.5-7B | 3,565 | 13.0% |
| Falcon3-7B-Base | 7,531 | 27.5% |
| Llama-3.1-8B-Instruct | 1,799 | 6.6% |
| Gemma-2-9B-IT | 802 | 2.9% |

Tabella: Belief Bank Facts

# Risultati: Falcon3 e Qwen2.5 su BeliefBankFacts

| Trainer → Tester | Approach | Type | AUROC (Tr) | AUROC (Te) |
|---|---|---|---|---|
| Qwen → Falcon | Baseline | attn | 0.999 | 0.992 |
| Qwen → Falcon | HApproach | attn | 0.999 | 0.984 |
| Qwen → Falcon | FullNonLinear | attn | 1.000 | 0.994 |
| Qwen → Falcon | ReducedNonLinear | attn | 0.999 | 0.995 |
| Qwen → Falcon | One-For-All | attn | 1.000 | 0.999 |

Tabella: Confronto Approcci (Qwen → Falcon)

- **One-For-All** e **FullNonLinear** ottengono i risultati migliori.
- **Baseline** è già molto forte in questo caso.

# Risultati: Llama-3.1 e Gemma-2 su BeliefBankFacts

| Trainer → Tester | Approach | Type | AUROC (Tr) | AUROC (Te) |
|---|---|---|---|---|
| Gemma → Llama | Baseline | attn | 0.986 | 0.974 |
| Gemma → Llama | HApproach | attn | 0.986 | 0.961 |
| Gemma → Llama | FullNonLinear | attn | 0.994 | 0.976 |
| Gemma → Llama | ReducedNonLinear | attn | 0.993 | 0.970 |
| Gemma → Llama | One-For-All | attn | 0.992 | 0.997 |

Tabella: Confronto Approcci (Gemma → Llama)

- **One-For-All** mostra un netto miglioramento rispetto alla Baseline.
- Gli approcci non lineari (Full/Reduced) non superano One-For-All.

# Risultati: Llama-3.1 e Gemma-2 su HaluEval

| Trainer → Tester | Approach | Type | AUROC (Tr) | AUROC (Te) |
|---|---|---|---|---|
| Gemma → Llama | Baseline | attn | 0.767 | 0.853 |
| Gemma → Llama | HApproach | attn | 0.767 | 0.811 |
| Gemma → Llama | FullNonLinear | attn | 0.790 | 0.834 |
| Gemma → Llama | ReducedNonLinear | attn | 0.772 | 0.788 |
| Gemma → Llama | One-For-All | attn | 0.768 | 0.855 |

Tabella: Confronto Approcci su HaluEval

- Su dataset complessi, **Baseline** rimane molto competitiva.
- **One-For-All** mantiene buone performance ma il gap è ridotto.

# Risultati

| Align Train | Target | Coppia | AUROC (Tr) | AUROC (Te) |
|---|---|---|---|---|
| BeliefBank | HaluEval | Gemma $\rightarrow$ Llama | 99.7% | 99.9% |
| BeliefBank | HaluEval | Llama $\rightarrow$ Gemma | 99.8% | 99.9% |
| HaluEval | BeliefBank | Gemma $\rightarrow$ Llama | 91.6% | 92.4% |
| HaluEval | BeliefBank | Llama $\rightarrow$ Gemma | 94.1% | 89.4% |

Tabella: Sintesi risultati Cross-Domain

**Osservazione**: Addestrare su task semplici (BeliefBank) generalizza meglio a task complessi (HaluEval) che viceversa.

# Discussione

- I metodi di trasferibilità sono generalmente efficaci.
- **One-For-All** emerge come il metodo più promettente e robusto.
- La trasferibilità è particolarmente alta per task di verifica fattuale semplice.

# Discussione

- **BeliefBank (Facts & Calibration)**:
  - Alta separabilità lineare (Acc $>$ 95%).
  - Concetto di verità "universale" e ben definito.
- **HaluEval**:
  - Molto più complesso (Acc 70-80%).
  - Le allucinazioni sono più specifiche del modello e del contesto.
  - Minore trasferibilità.

# Discussione

- **Falcon3 & Qwen2.5**: Altissima compatibilità, suggerisce similarità strutturale.
- **Llama-3.1 & Gemma-2**: Buona compatibilità, ma minore su task complessi.
- **Layer**:
  - *Attention* e *Hidden States* spesso più stabili di *MLP*.
  - One-For-All è robusto su tutte le componenti.

# Discussione

- **Simple** $\rightarrow$ **Complex**: Funziona molto bene.
- **Complex** $\rightarrow$ **Simple**: Funziona peggio.
- **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.
- **Applicazioni Real-Time**: Implementare il prober in sistemi di monitoraggio live.

# Grazie per l'attenzione! 🚀